Minería de datos y Patrones

Version 2024-I

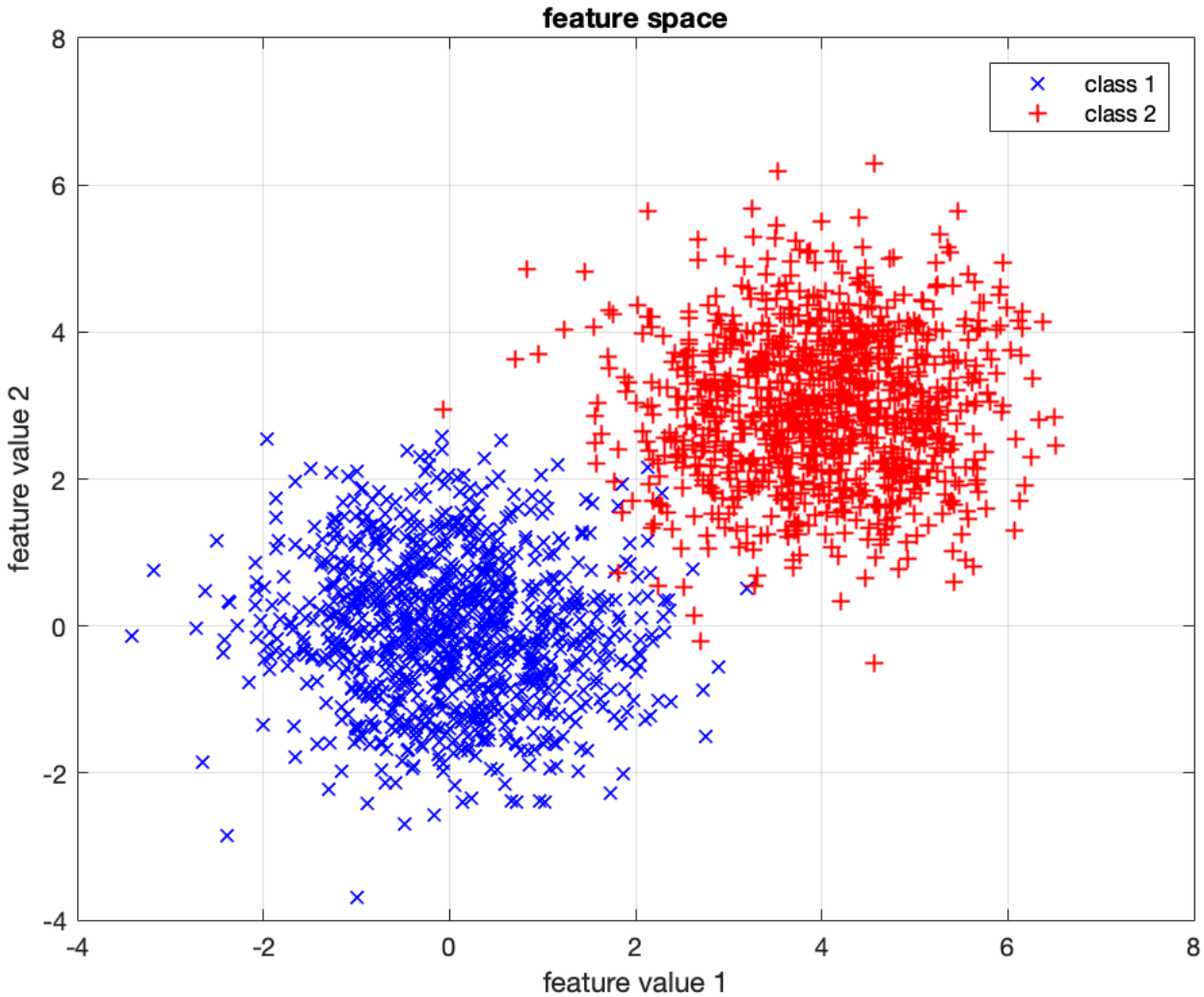
Árboles de Decisión

[Capítulo 4]

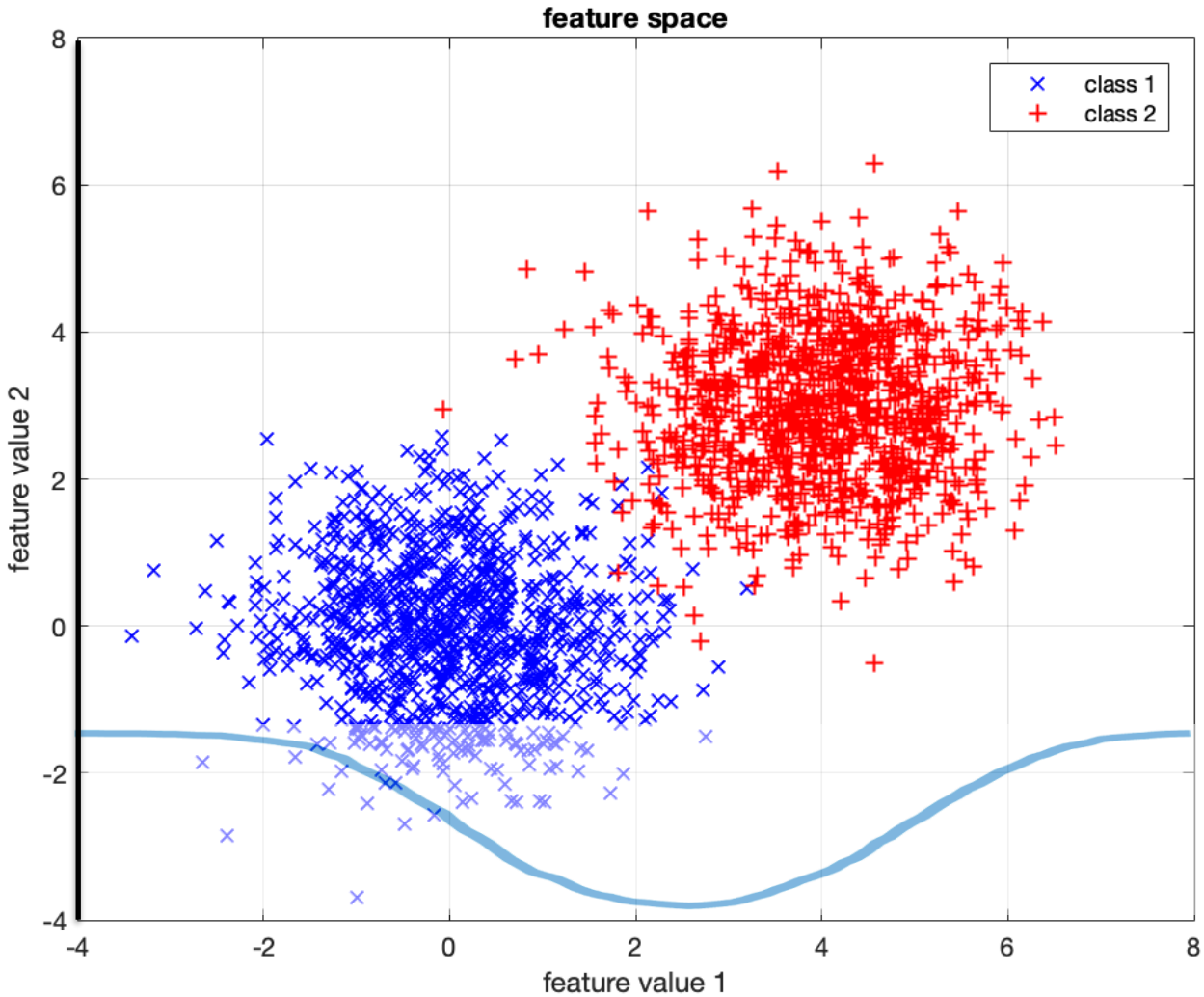
Dr. José Ramón Iglesias
DSP-ASIC BUILDER GROUP
Director Semillero TRIAC
Ingenieria Electronica
Universidad Popular del Cesar

Árboles de Decisión

Árboles de Decisión (Training data)

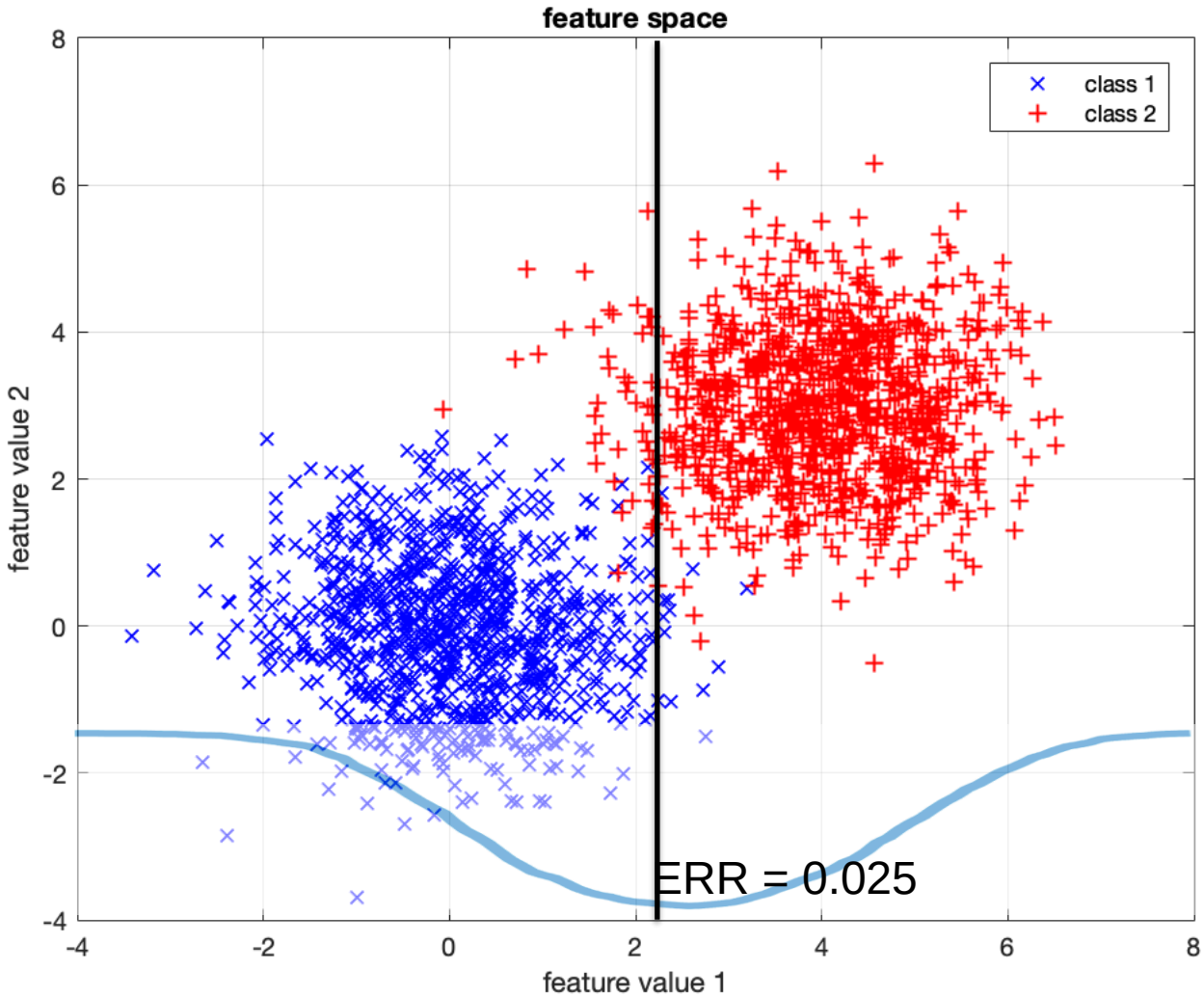


Árboles de Decisión (Training data)



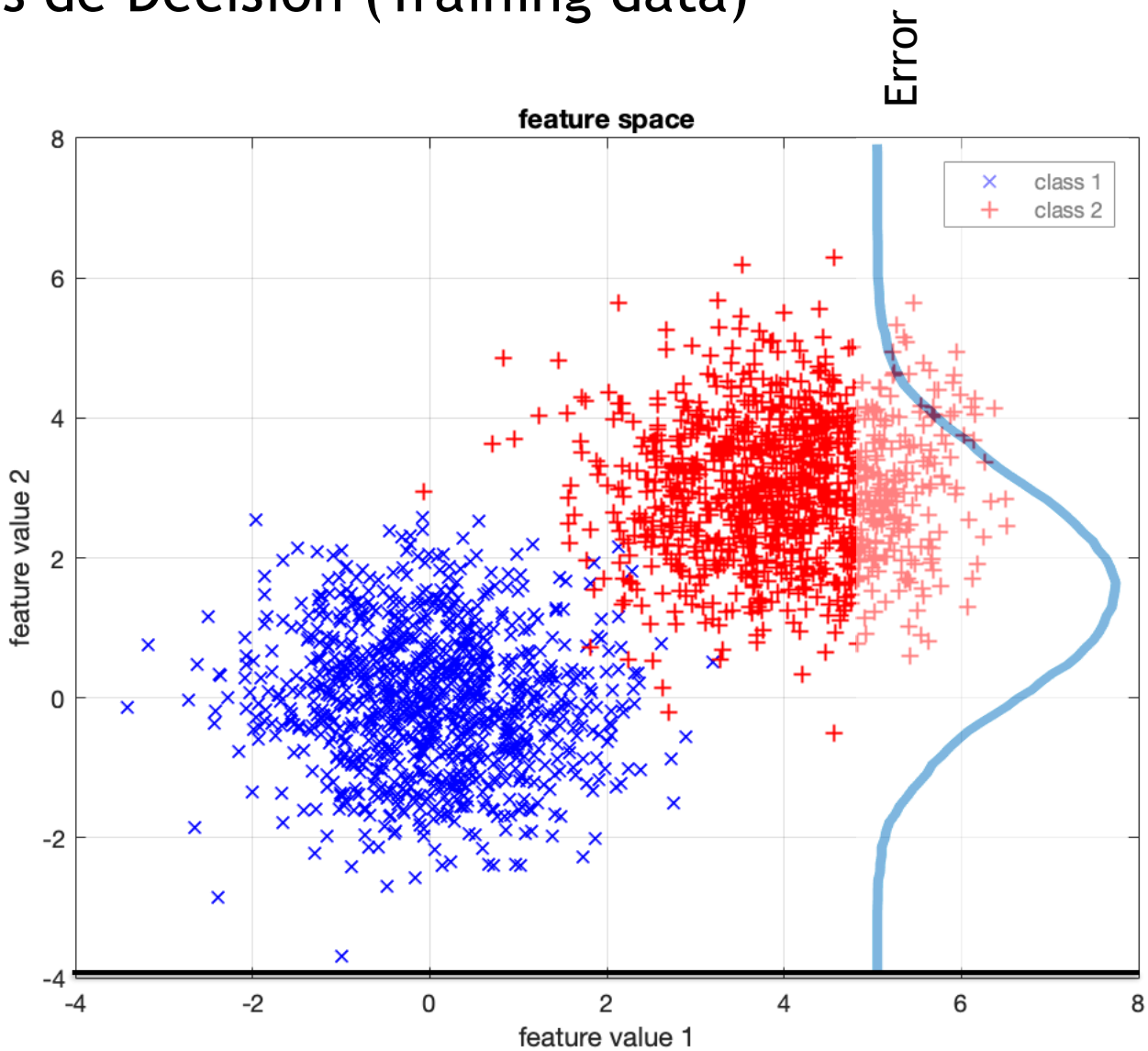
Error

Árboles de Decisión (Training data)

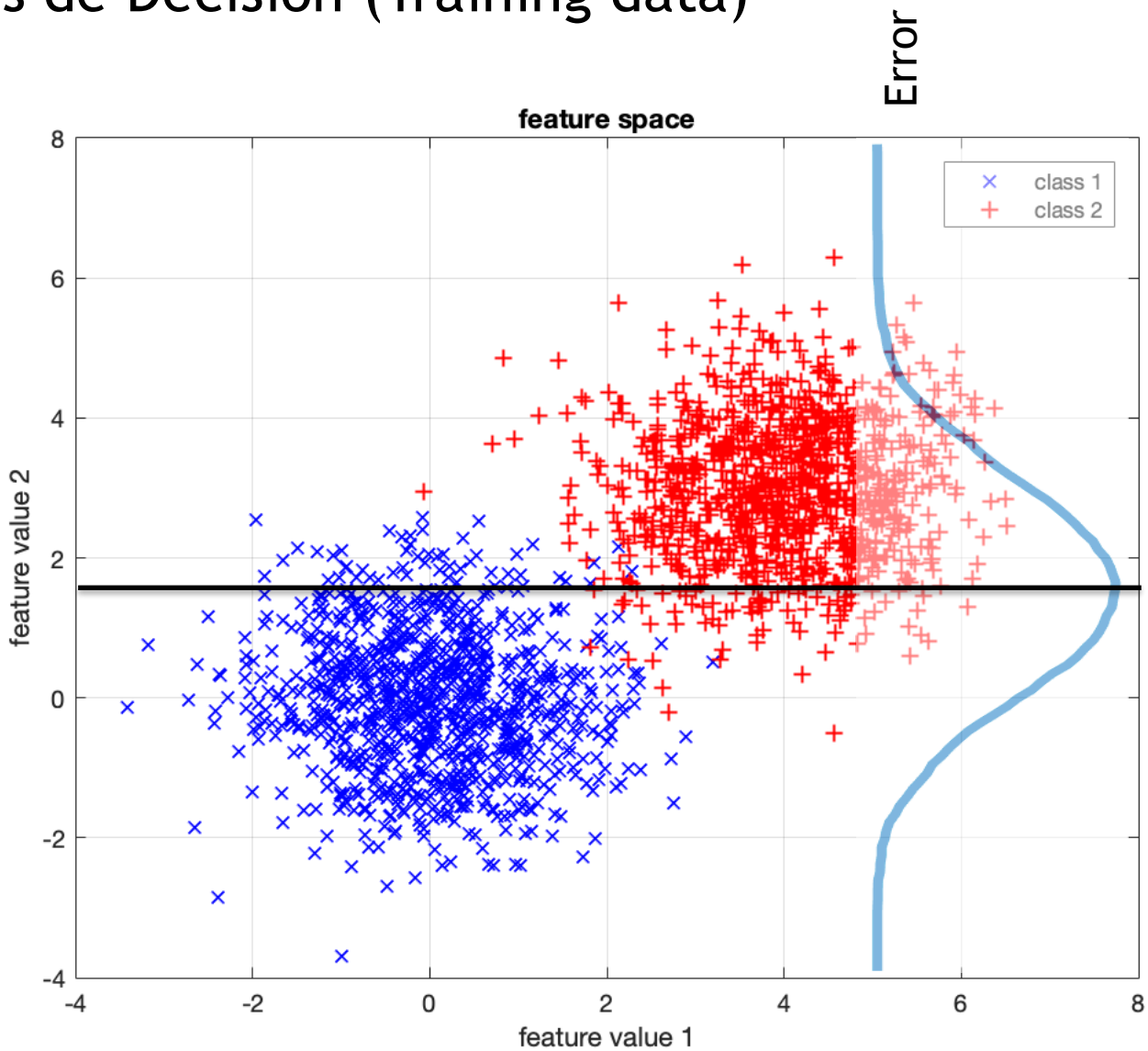


Error

Árboles de Decisión (Training data)

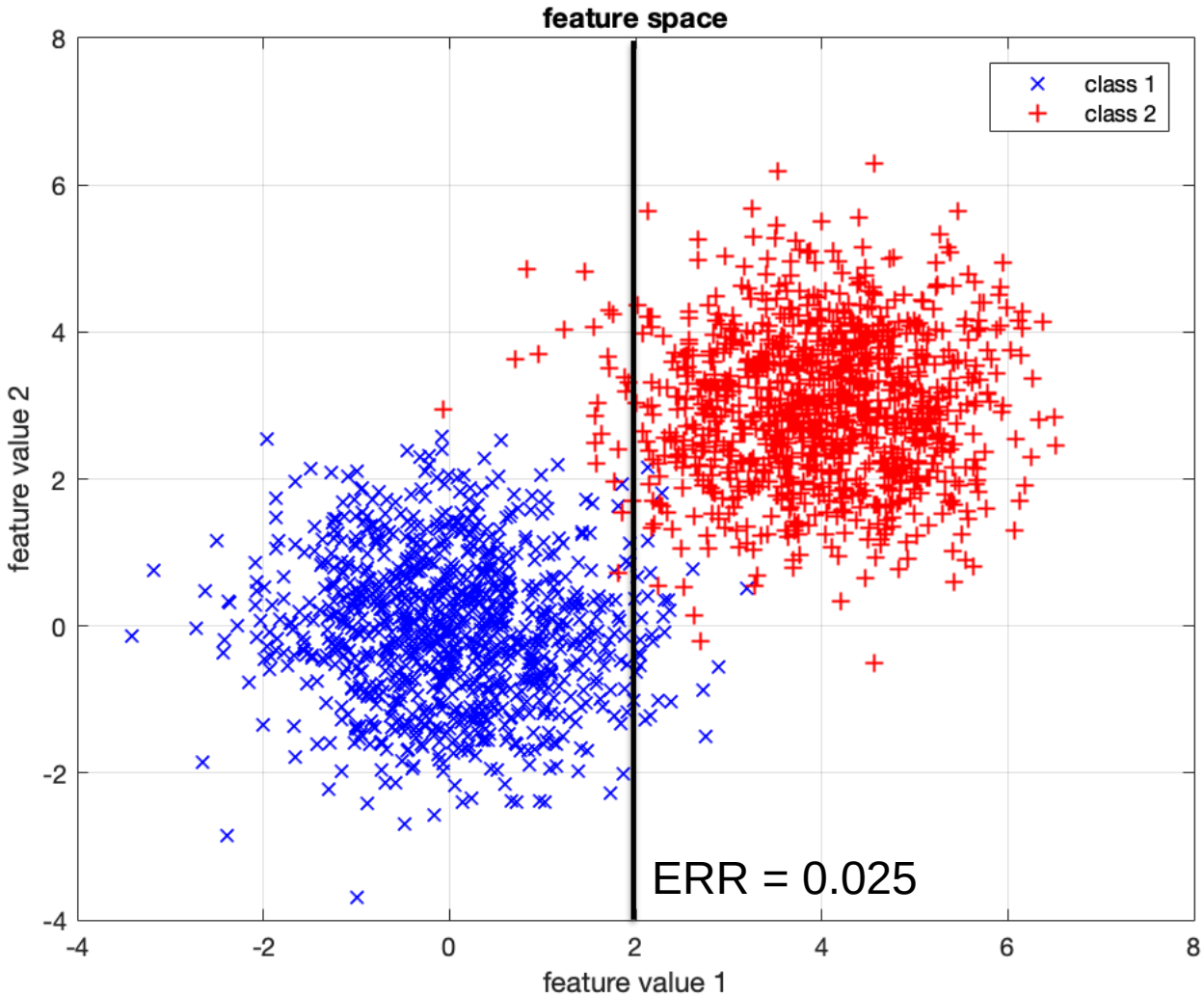


Árboles de Decisión (Training data)

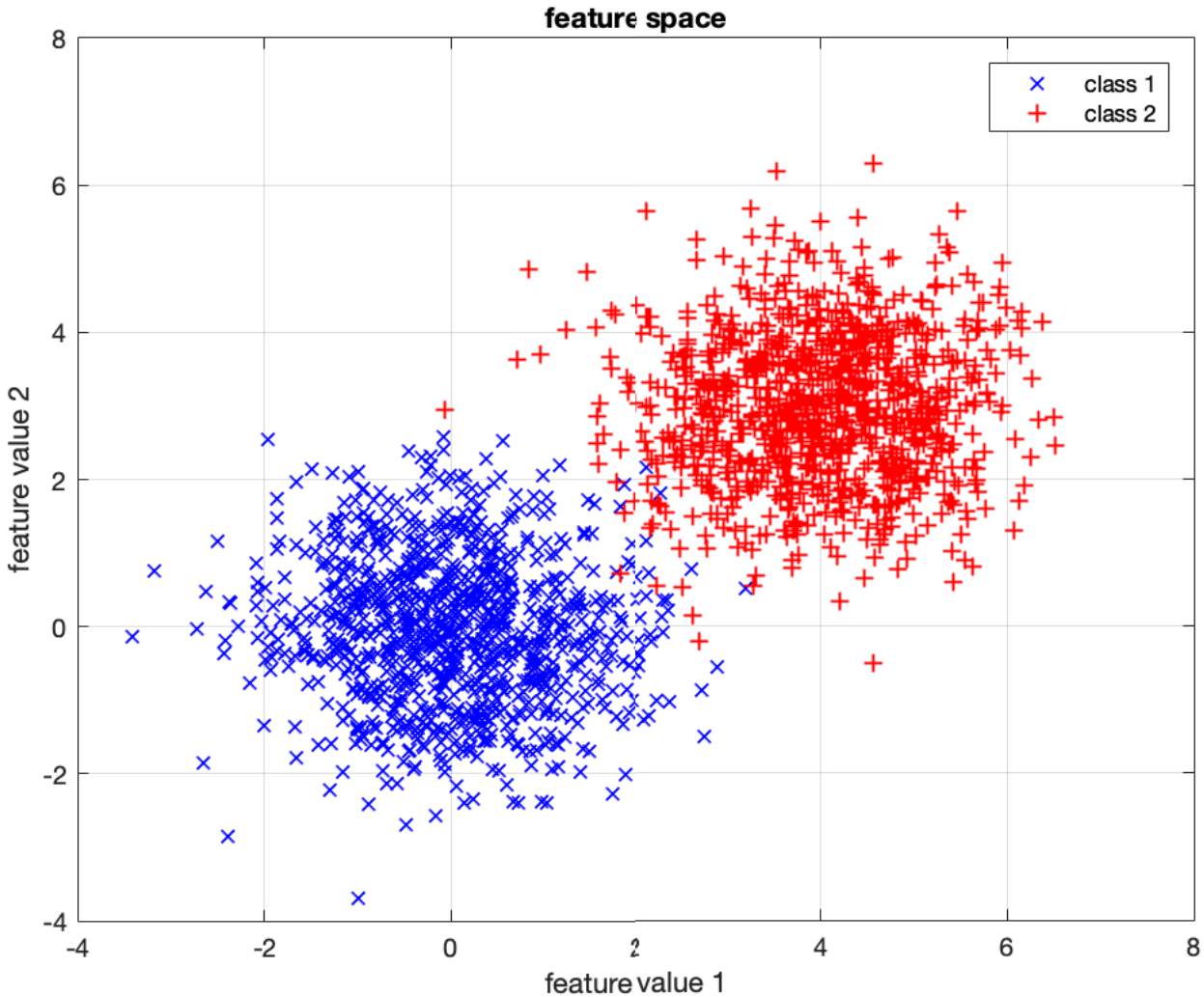


ERR = 0.066

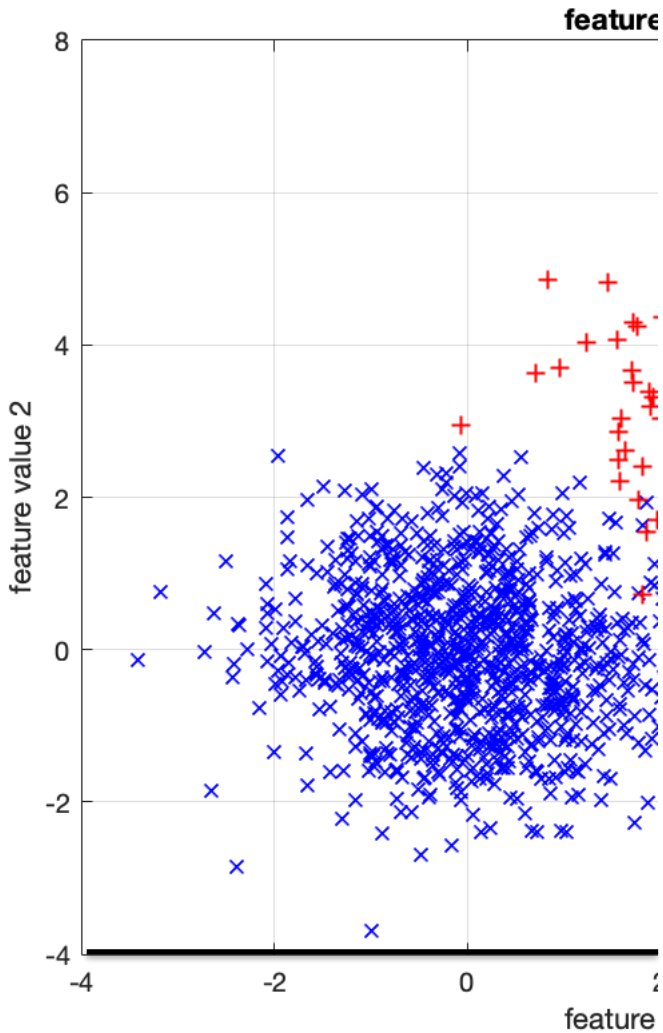
Árboles de Decisión (Training data)

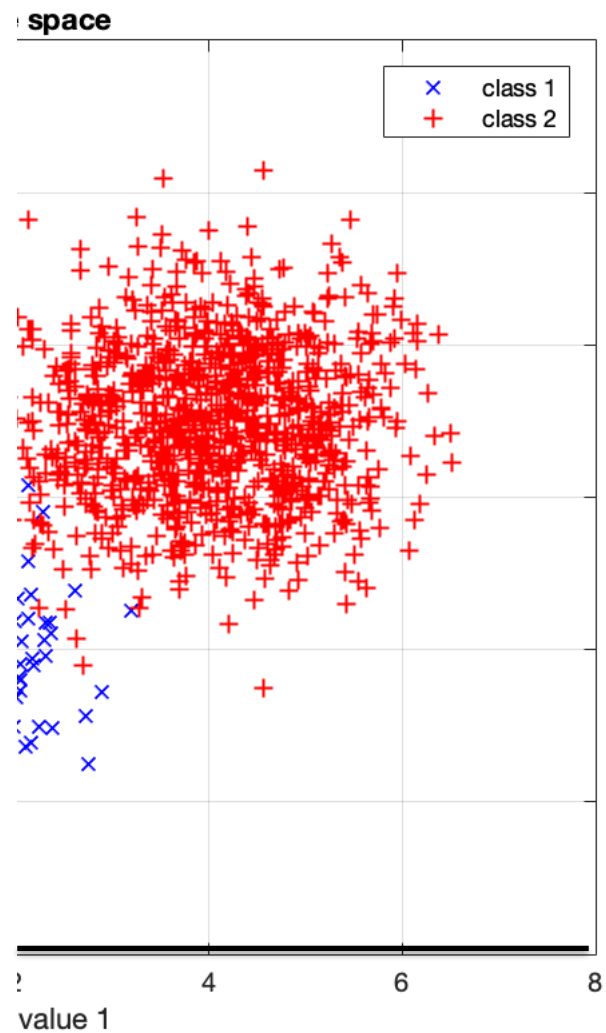


Árboles de Decisión (Training data)

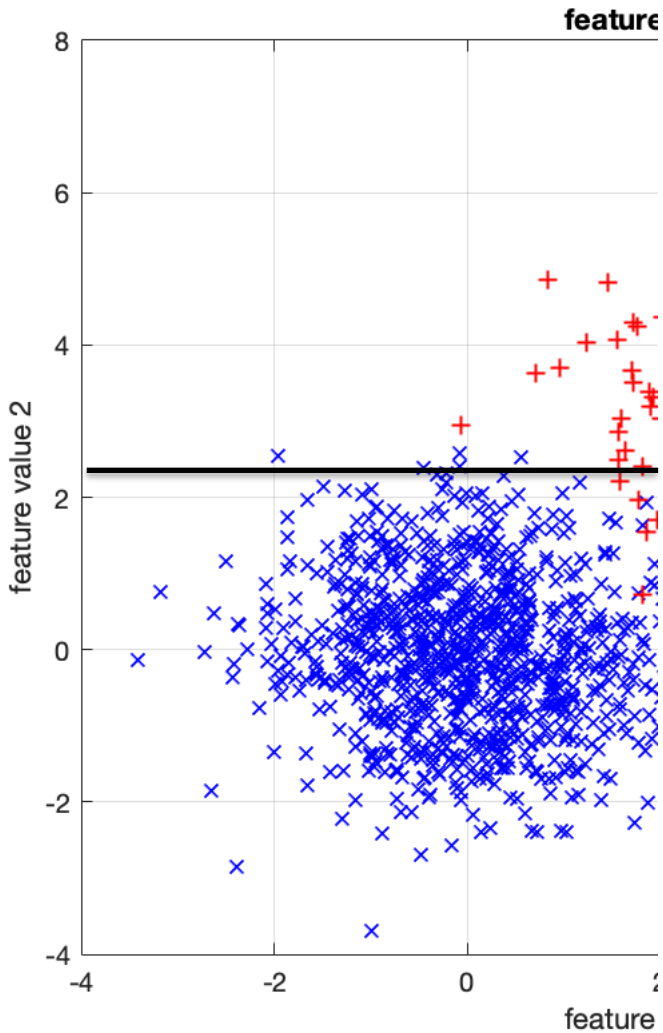


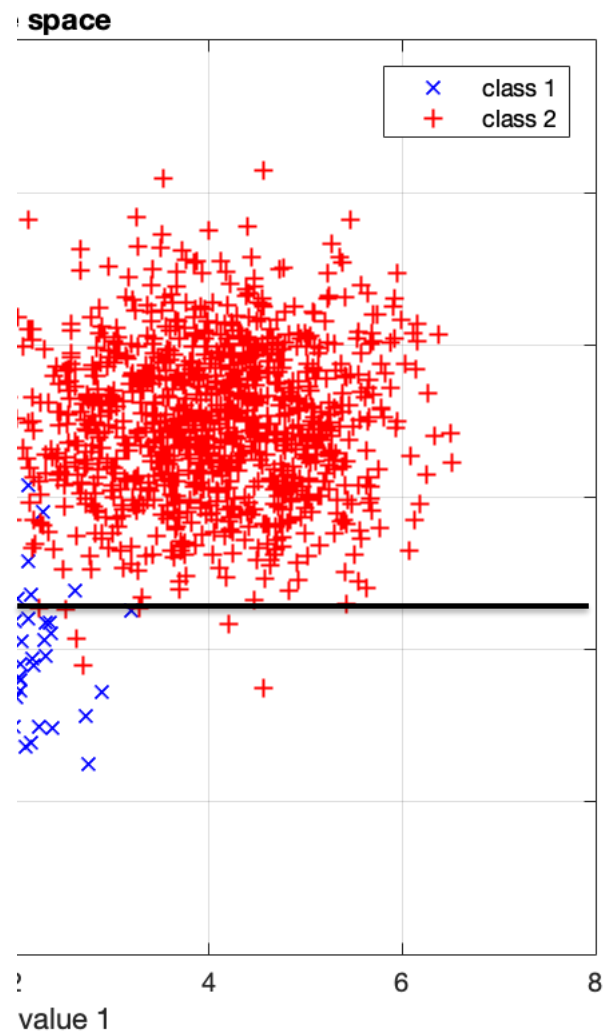
Árboles de Decisión (Training data)



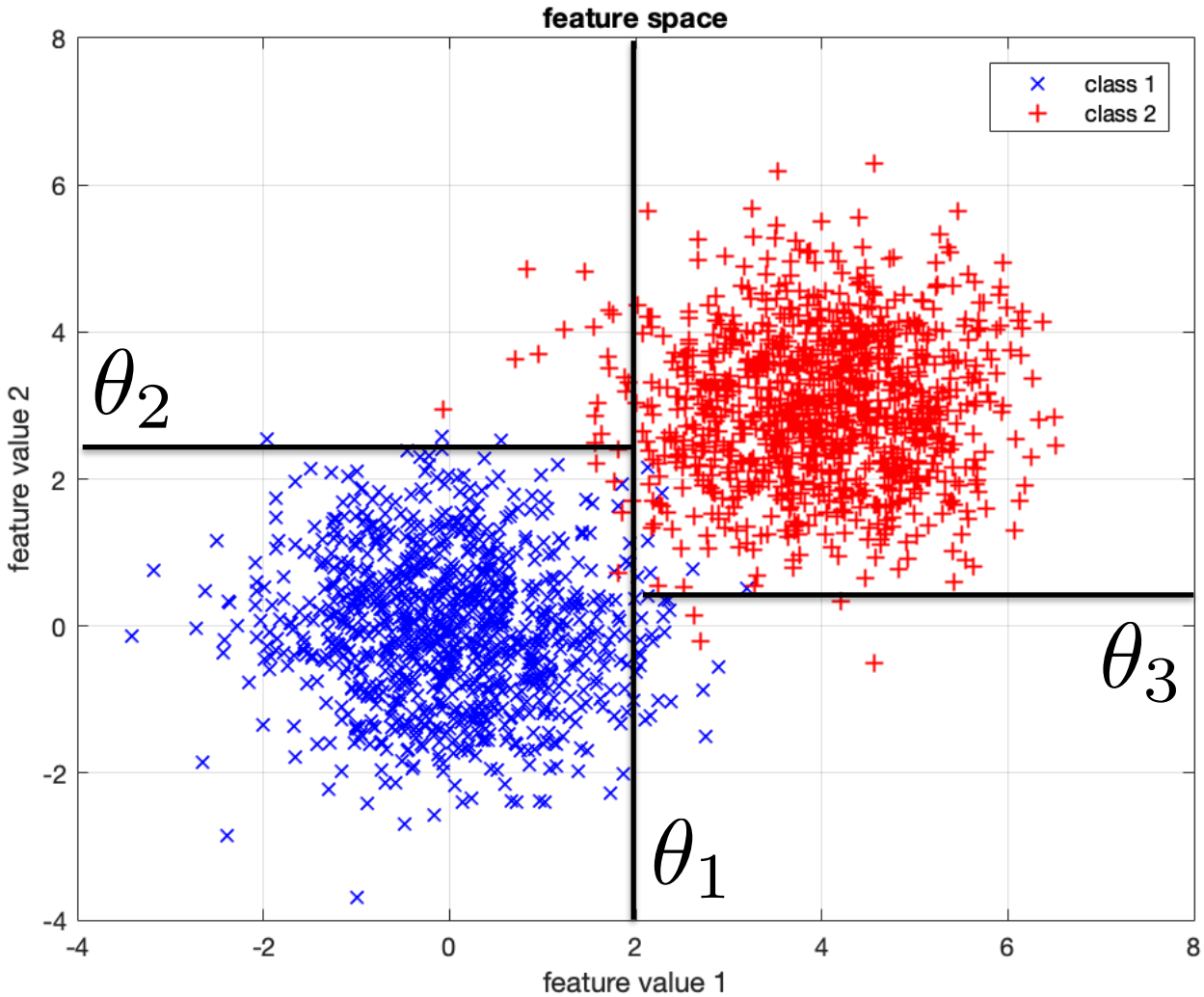


Árboles de Decisión (Training data)



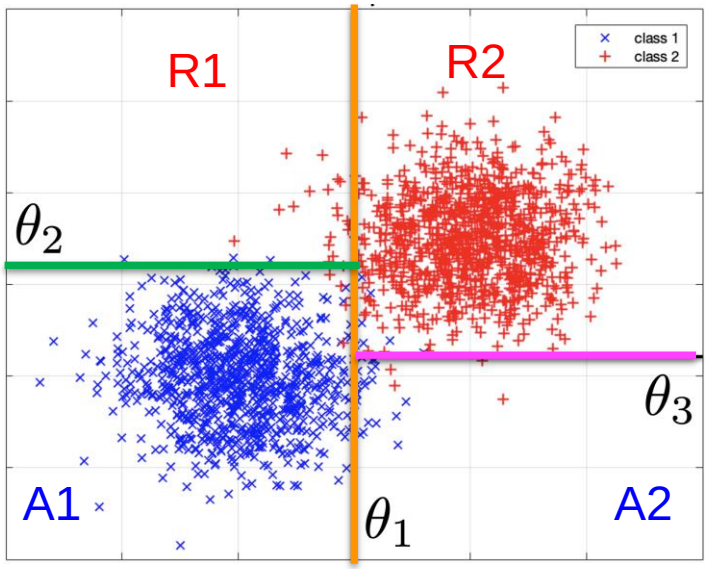


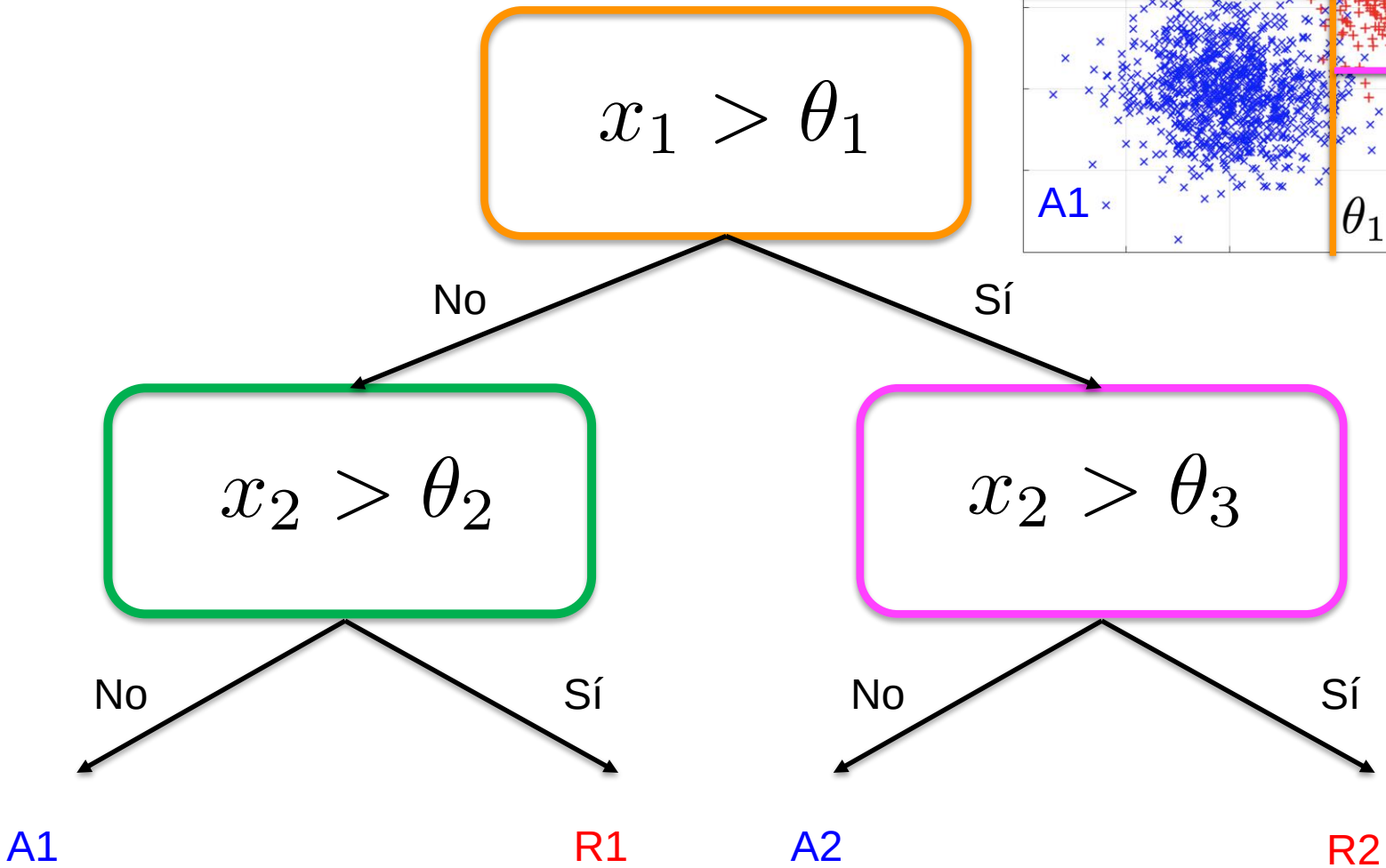
Árboles de Decisión (Training data)



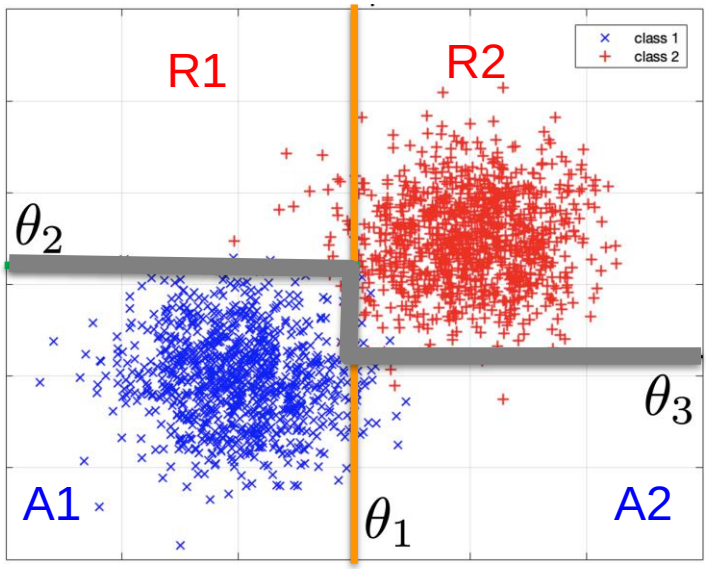
etc...

Árbol


class 1
class 2
R1
R2
θ_2
θ_3
A1
θ_1
A2


$x_1 > \theta_1$
No
Sí
$x_2 > \theta_2$
$x_2 > \theta_3$
No
Sí
No
Sí
A1
R1
A2
R2

Árbol



$$x_1 > \theta_1$$

No

Sí

$$x_2 > \theta_2$$

$$x_2 > \theta_3$$

No

Sí

No

Sí

A1

R1

A2

R2

Métricas usadas para el error:

- Error de clasificación 1 - Accuracy

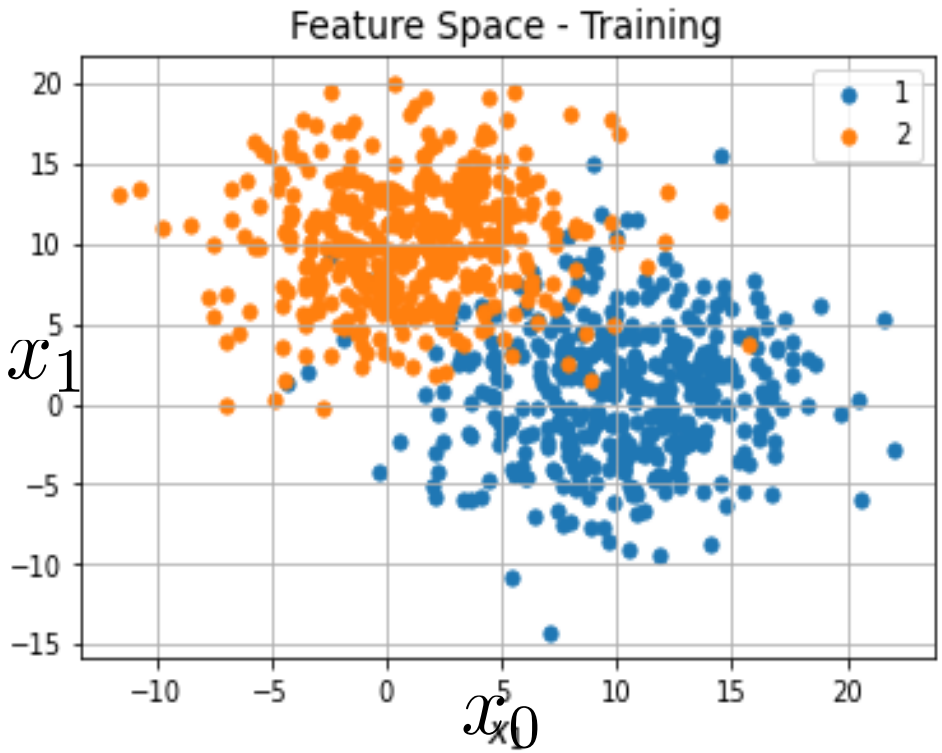
- Entropía $$-\sum_{k=1}^{K} p_k \log_2(p_k)$$

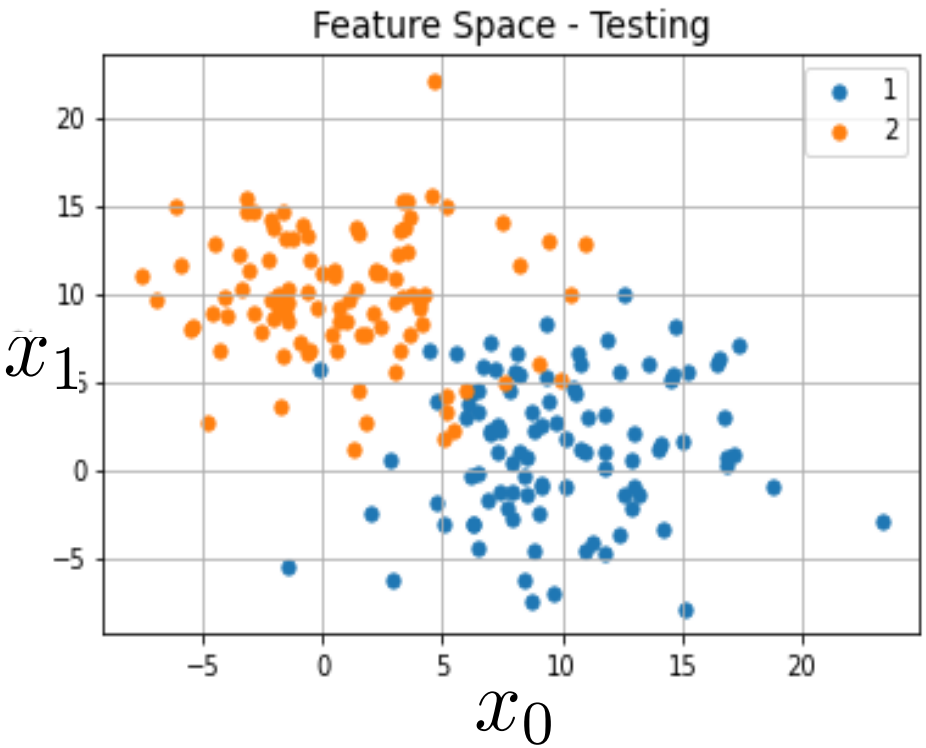
- Índice Gini $$\sum_{k=1}^{K} p_k(1-p_k)$$

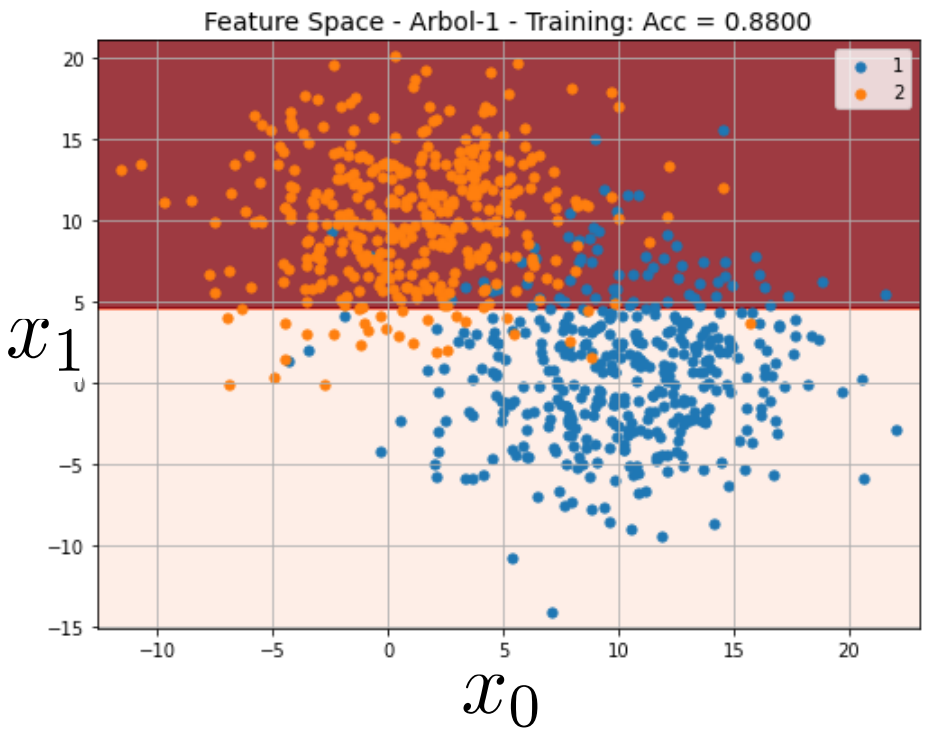
p_k Probabilidad de clasificar bien la clase *k*

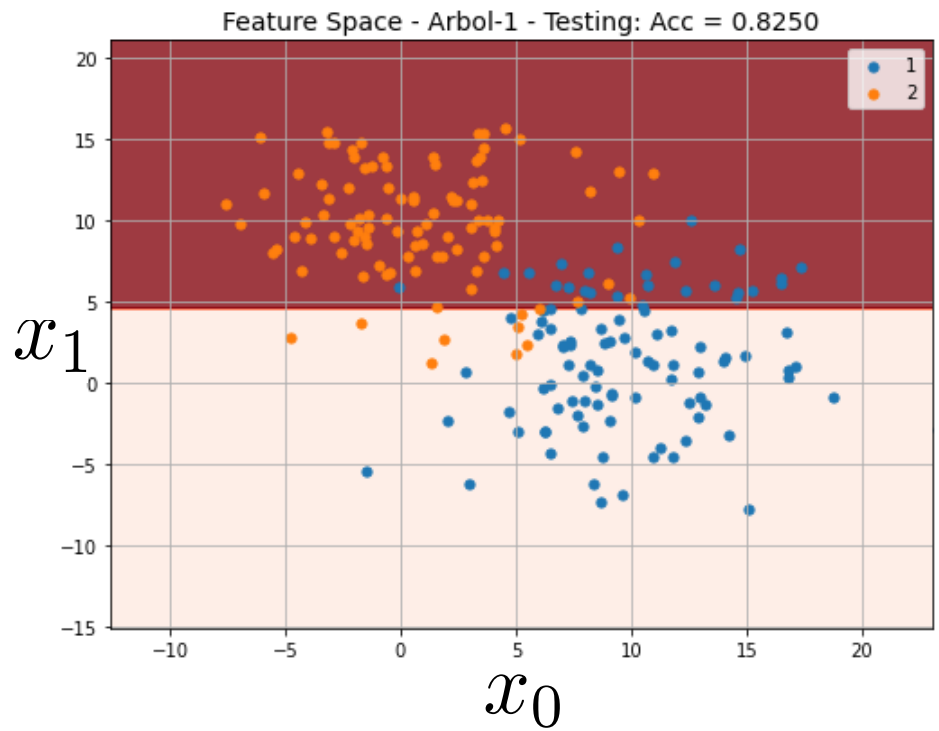
EJEMPLO

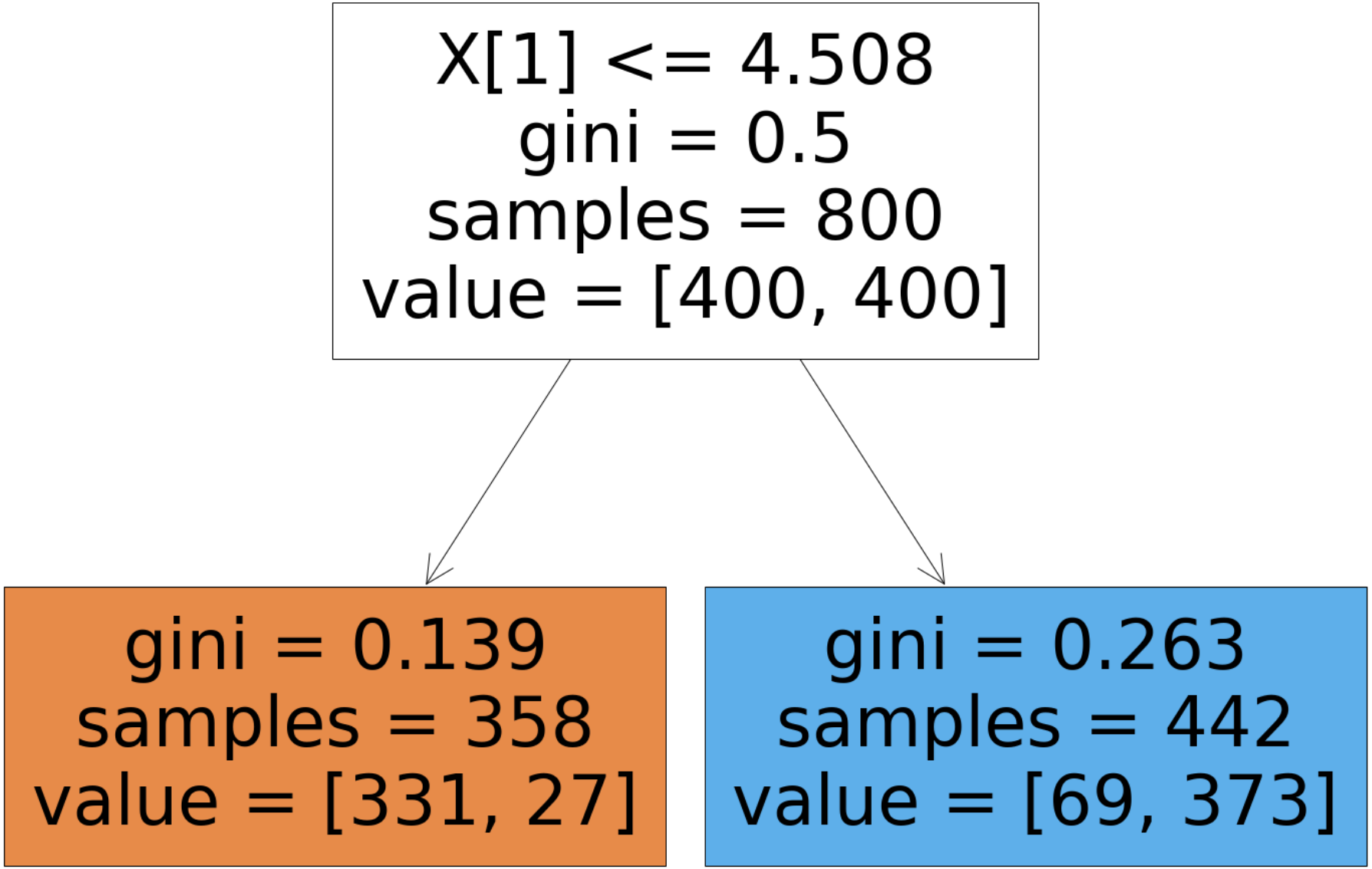
Datos Training/Testing

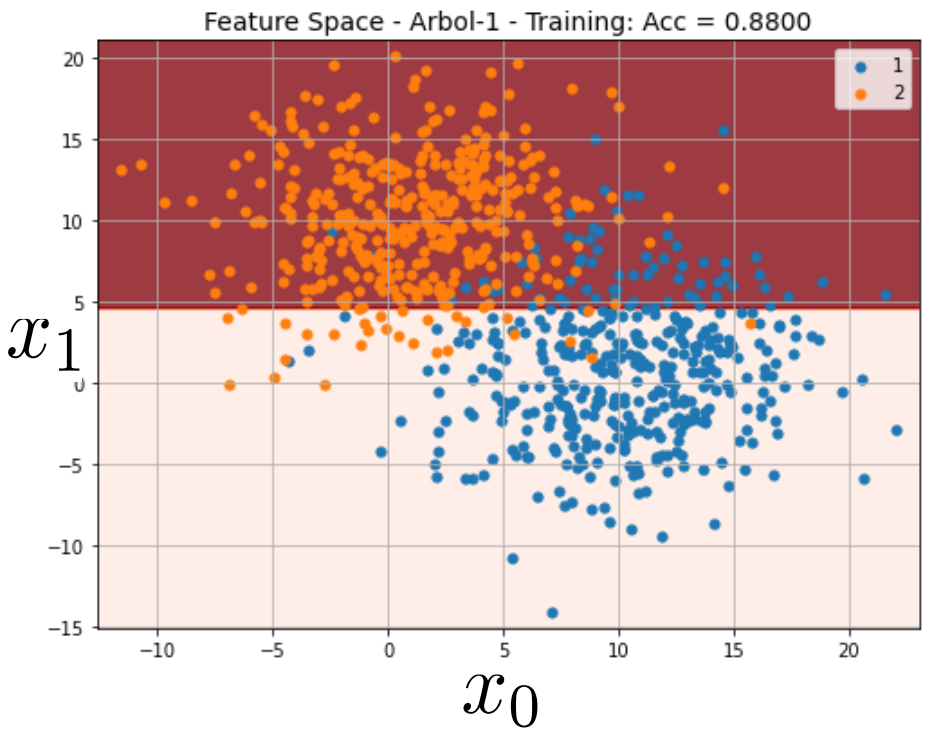


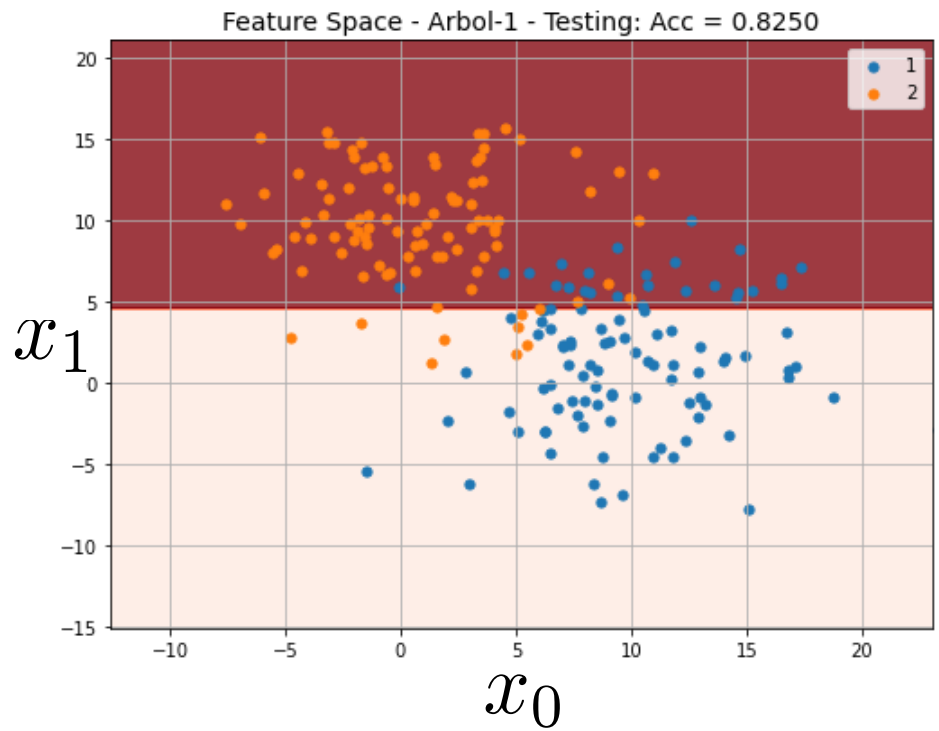


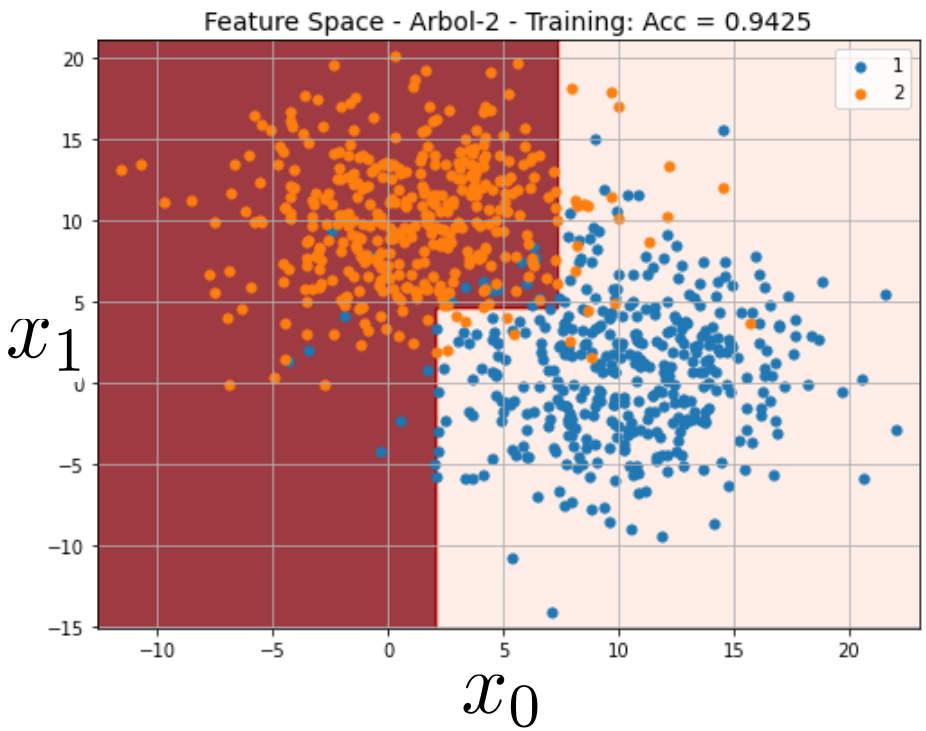

Feature Space - Arbol-1 - Training: Acc = 0.8800
1
2
x_1
x_0

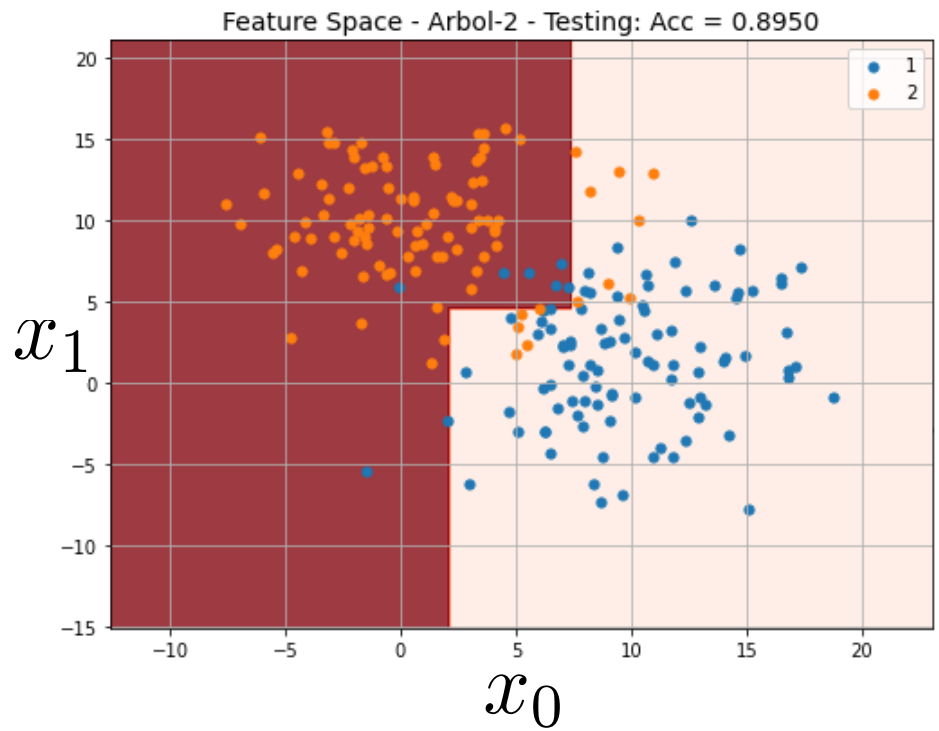

Feature Space - Arbol-1 - Testing: Acc = 0.8250
1
2
x_1
x_0

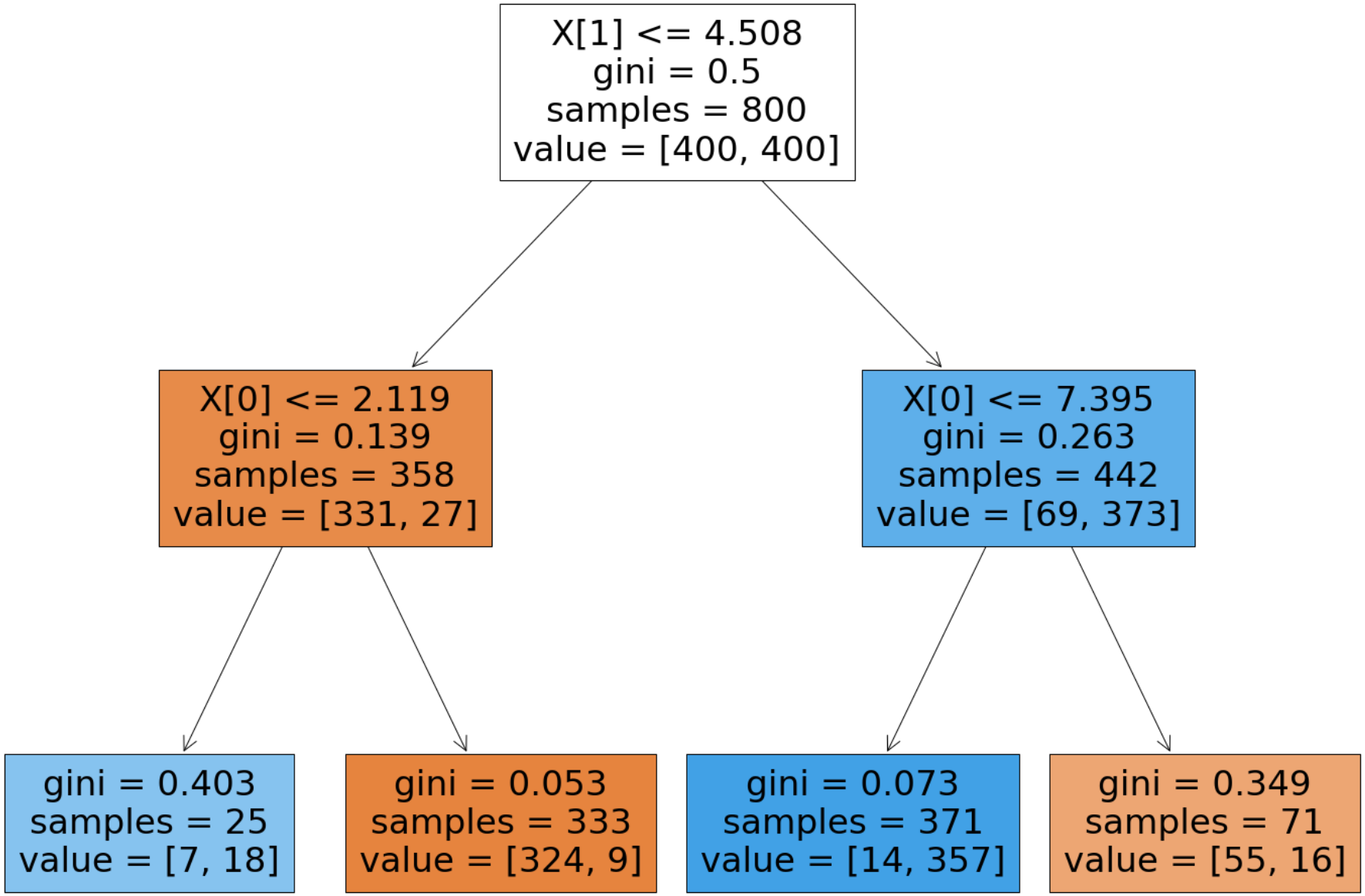

X[1] <= 4.508
gini = 0.5
samples = 800
value = [400, 400]
gini = 0.139
samples = 358
value = [331, 27]
gini = 0.263
samples = 442
value = [69, 373]

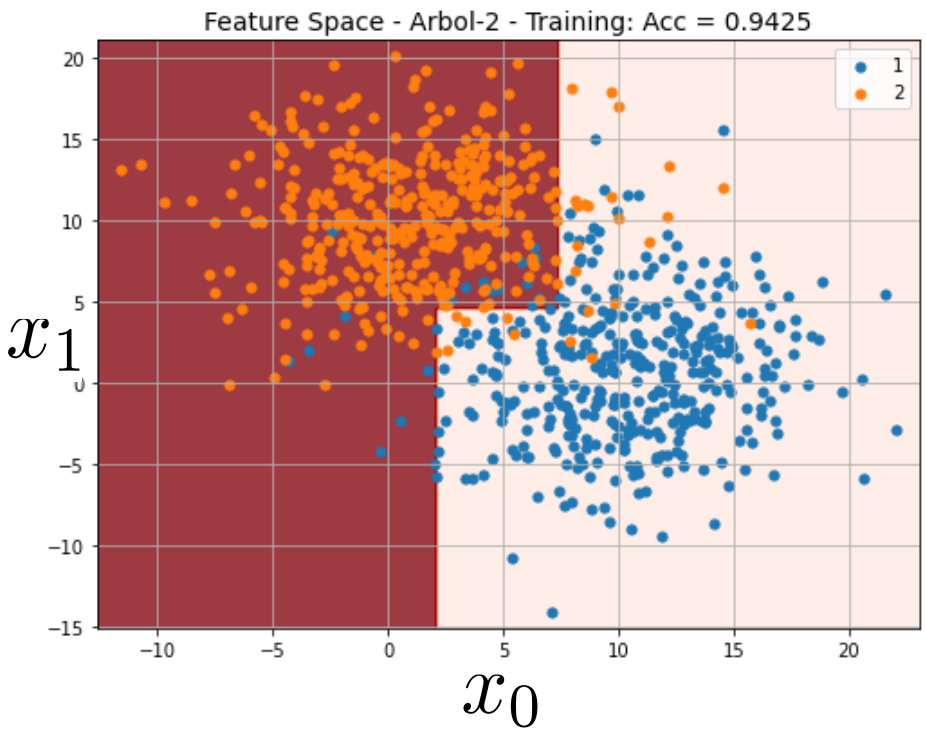

Feature Space - Arbol-1 - Training: Acc = 0.8800
1
2
x_1
x_0

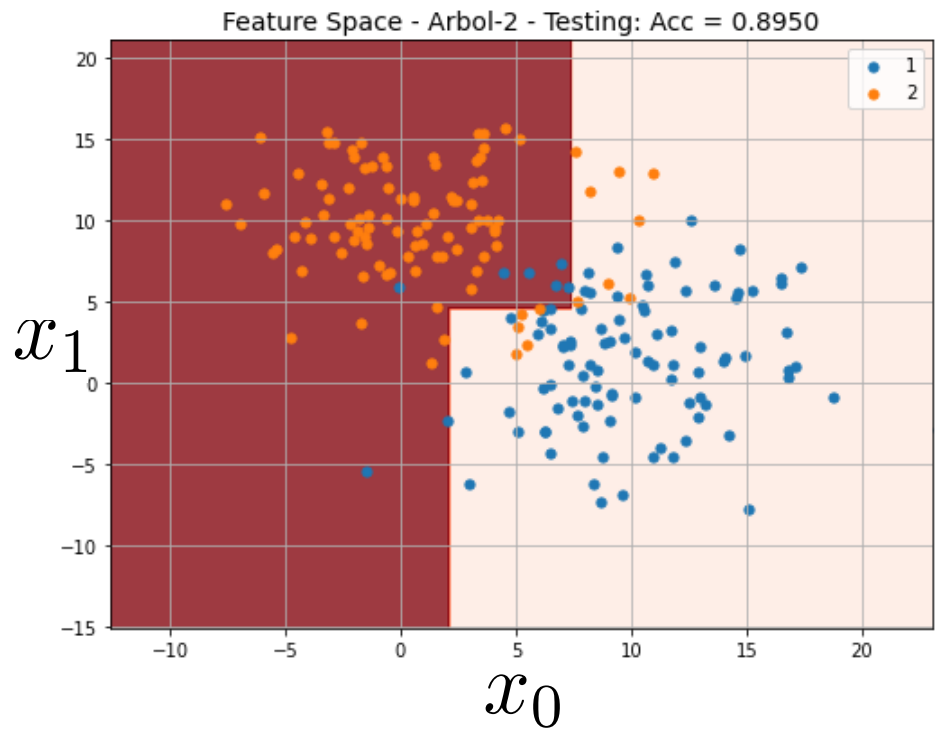

Feature Space - Arbol-1 - Testing: Acc = 0.8250
1
2
x_1
x_0

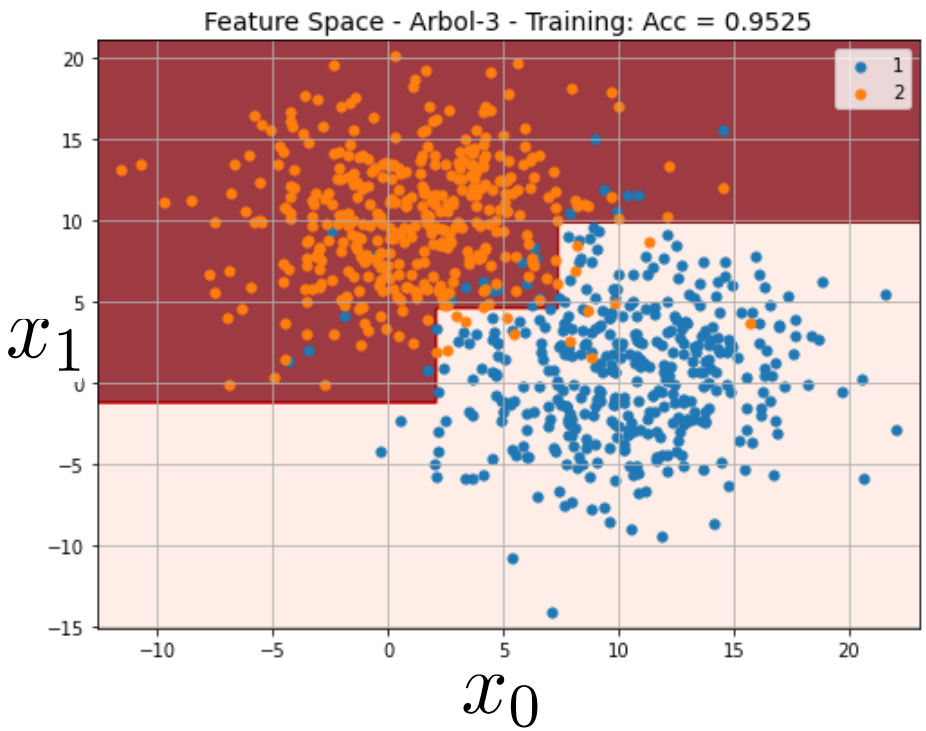

Feature Space - Arbol-2 - Training: Acc = 0.9425
1
2
x_1
x_0

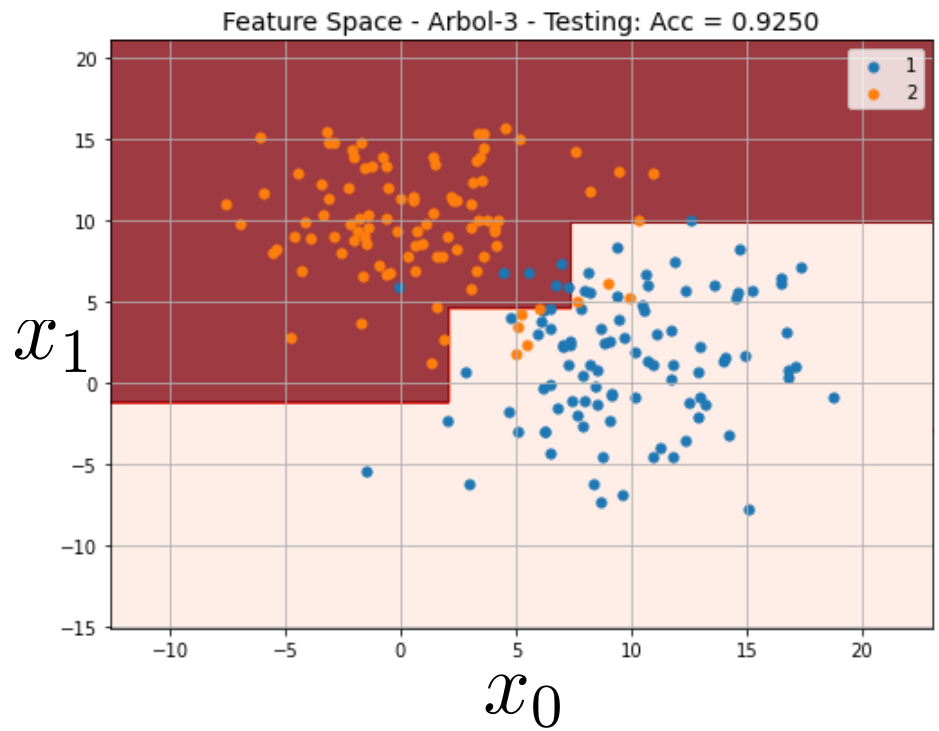

Feature Space - Arbol-2 - Testing: Acc = 0.8950
1
2
x_1
x_0

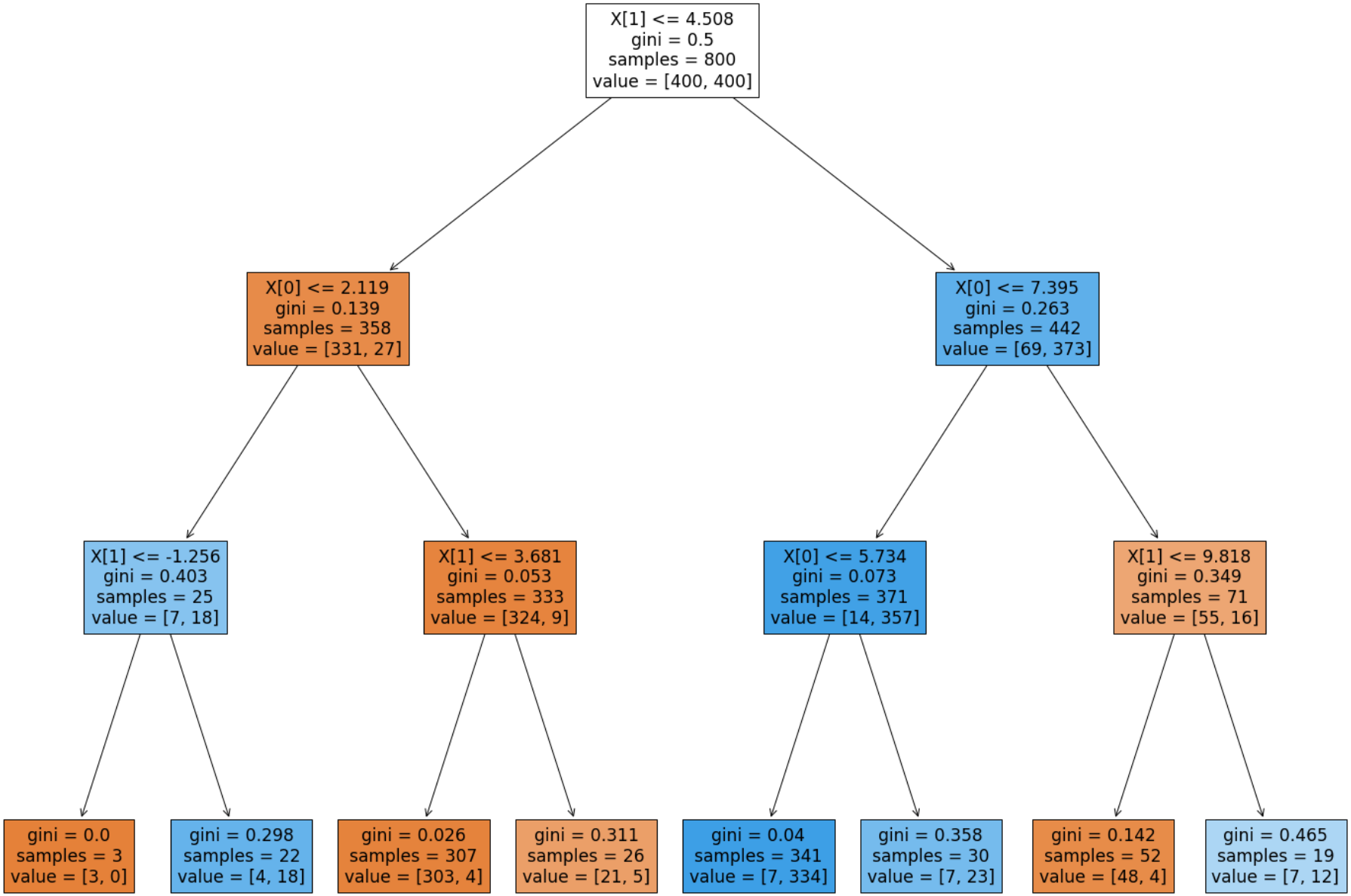

X[1] <= 4.508
gini = 0.5
samples = 800
value = [400, 400]
X[0] <= 2.119
gini = 0.139
samples = 358
value = [331, 27]
X[0] <= 7.395
gini = 0.263
samples = 442
value = [69, 373]
gini = 0.403
samples = 25
value = [7, 18]
gini = 0.053
samples = 333
value = [324, 9]
gini = 0.073
samples = 371
value = [14, 357]
gini = 0.349
samples = 71
value = [55, 16]

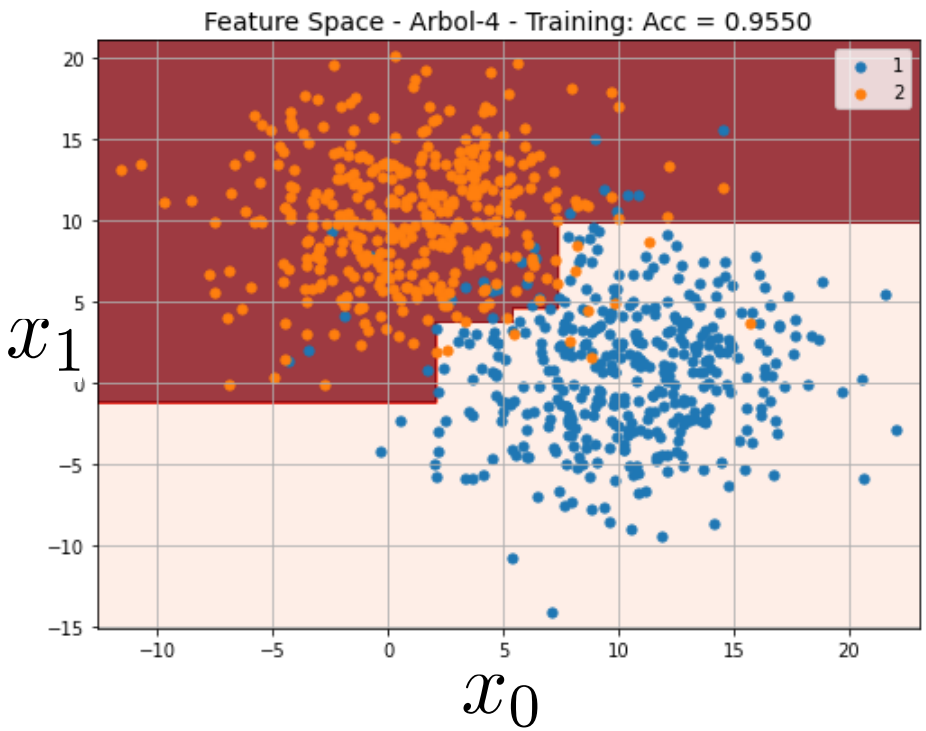

Feature Space - Arbol-2 - Training: Acc = 0.9425
1
2
x_1
x_0

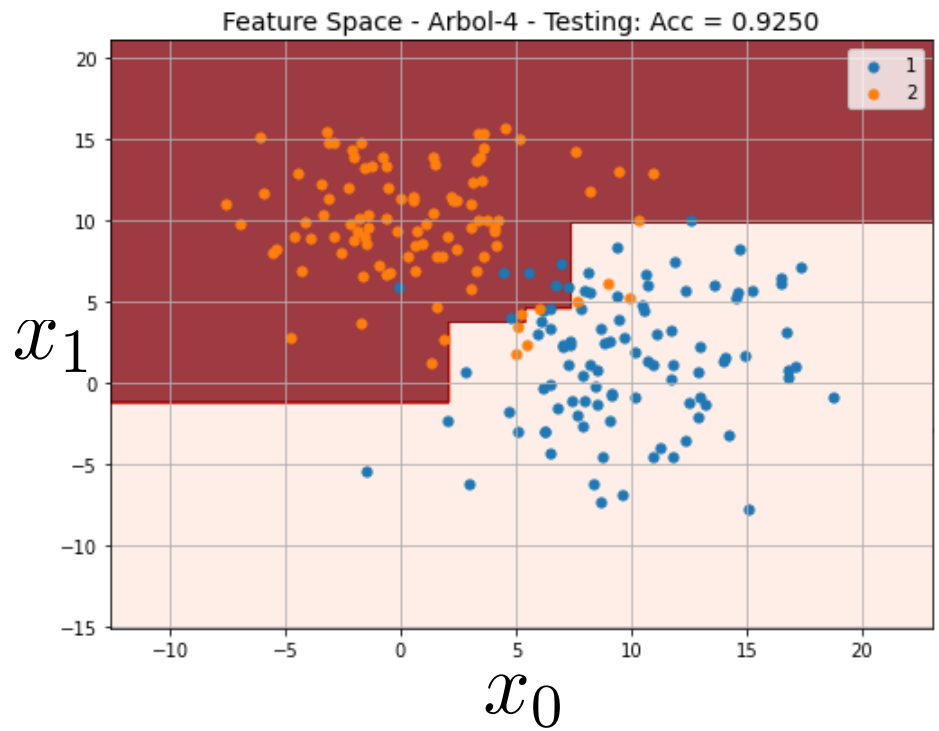

Feature Space - Arbol-2 - Testing: Acc = 0.8950
1
2
x_1
x_0

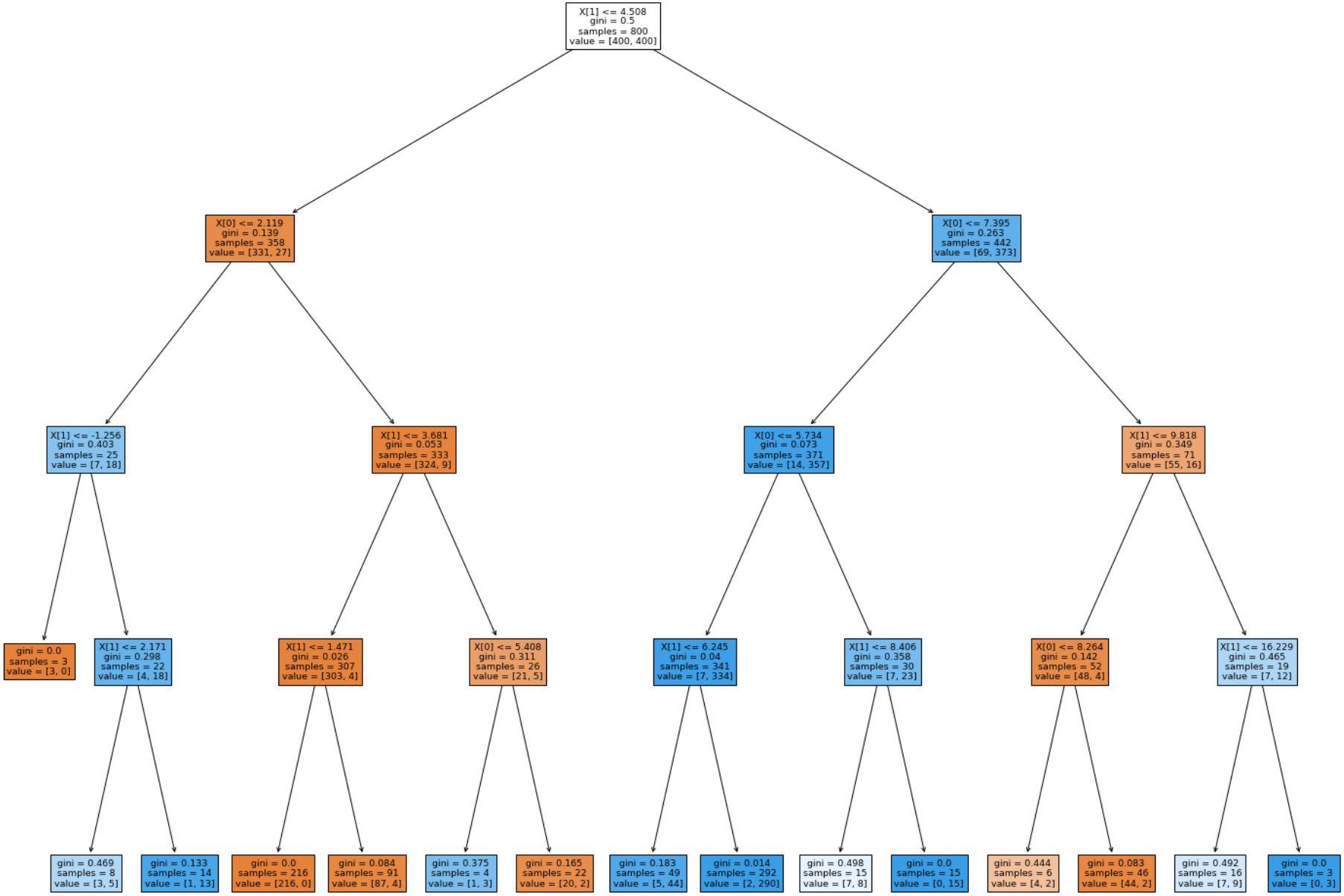

Feature Space - Arbol-3 - Training: Acc = 0.9525
1
2
x_1
x_0

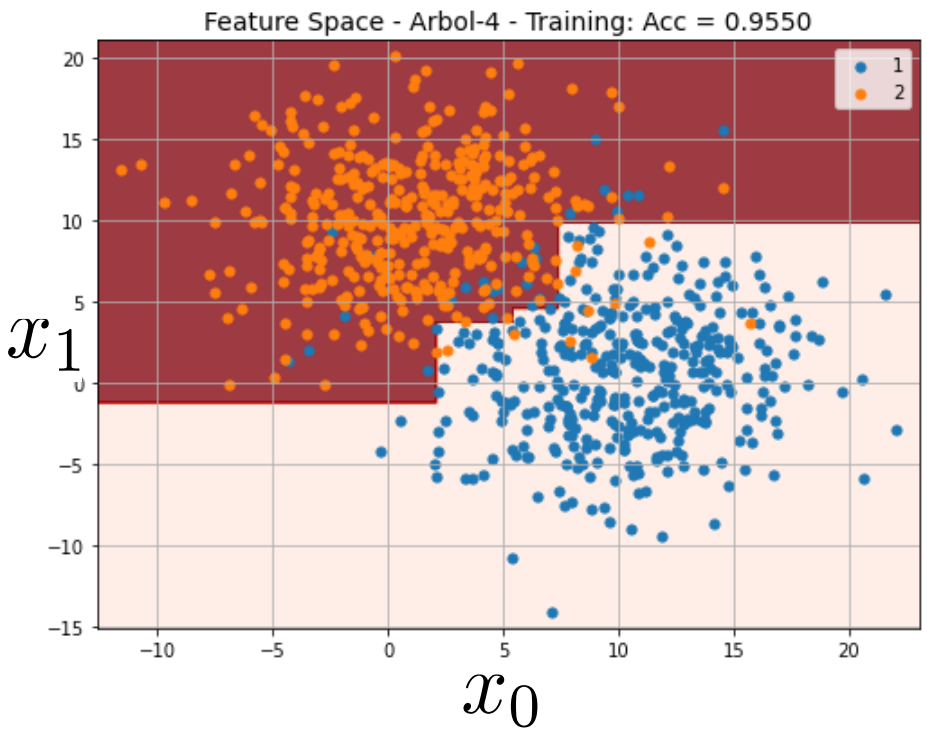

Feature Space - Arbol-3 - Testing: Acc = 0.9250
1
2
x_1
x_0

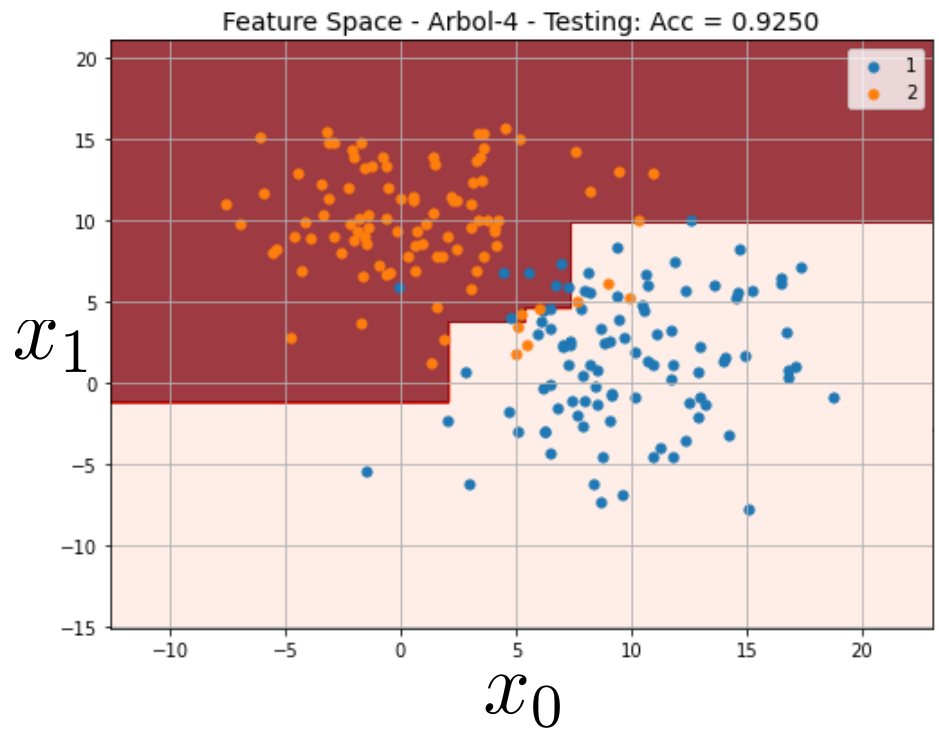

X[1] <= 4.508
gini = 0.5
samples = 800
value = [400, 400]
X[0] <= 2.119
gini = 0.139
samples = 358
value = [331, 27]
X[0] <= 7.395
gini = 0.263
samples = 442
value = [69, 373]
X[1] <= -1.256
gini = 0.403
samples = 25
value = [7, 18]
X[1] <= 3.681
gini = 0.053
samples = 333
value = [324, 9]
X[0] <= 5.734
gini = 0.073
samples = 371
value = [14, 357]
X[1] <= 9.818
gini = 0.349
samples = 71
value = [55, 16]
gini = 0.0
samples = 3
value = [3, 0]
gini = 0.298
samples = 22
value = [4, 18]
gini = 0.026
samples = 307
value = [303, 4]
gini = 0.311
samples = 26
value = [21, 5]
gini = 0.04
samples = 341
value = [7, 334]
gini = 0.358
samples = 30
value = [7, 23]
gini = 0.142
samples = 52
value = [48, 4]
gini = 0.465
samples = 19
value = [7, 12]

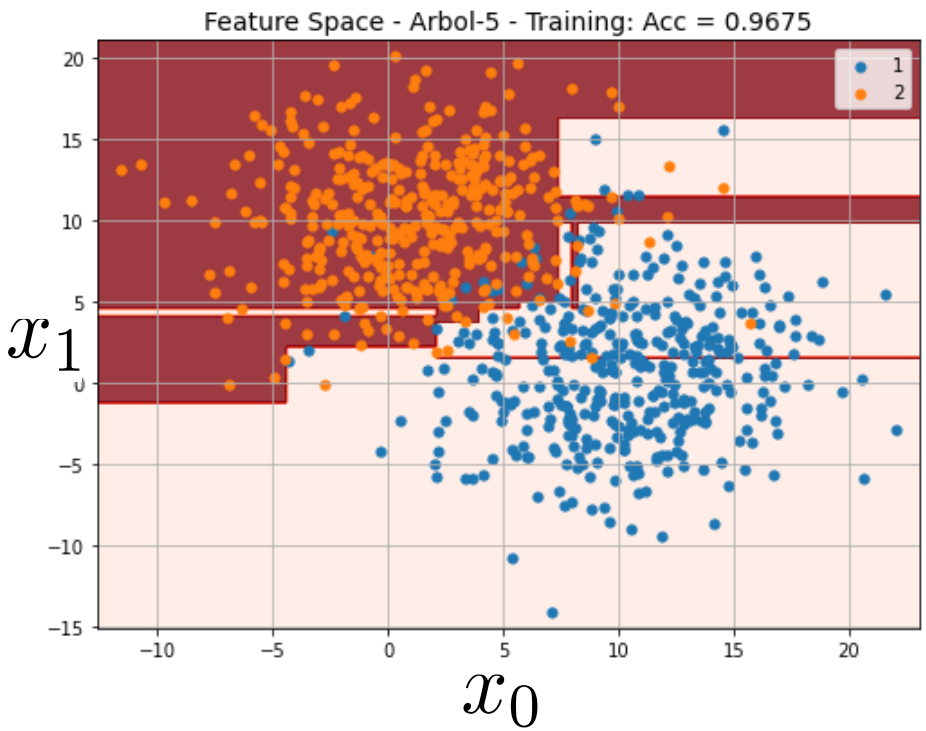

Feature Space - Arbol-4 - Training: Acc = 0.9550
1
2
x_1
x_0

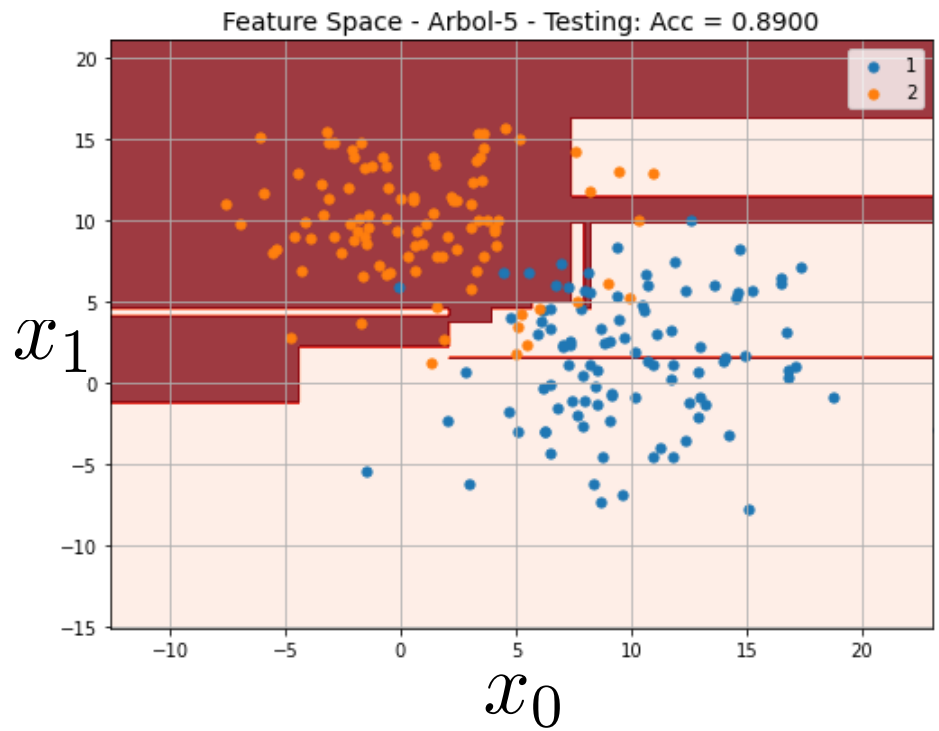

Feature Space - Arbol-4 - Testing: Acc = 0.9250
1
2
x_1
x_0

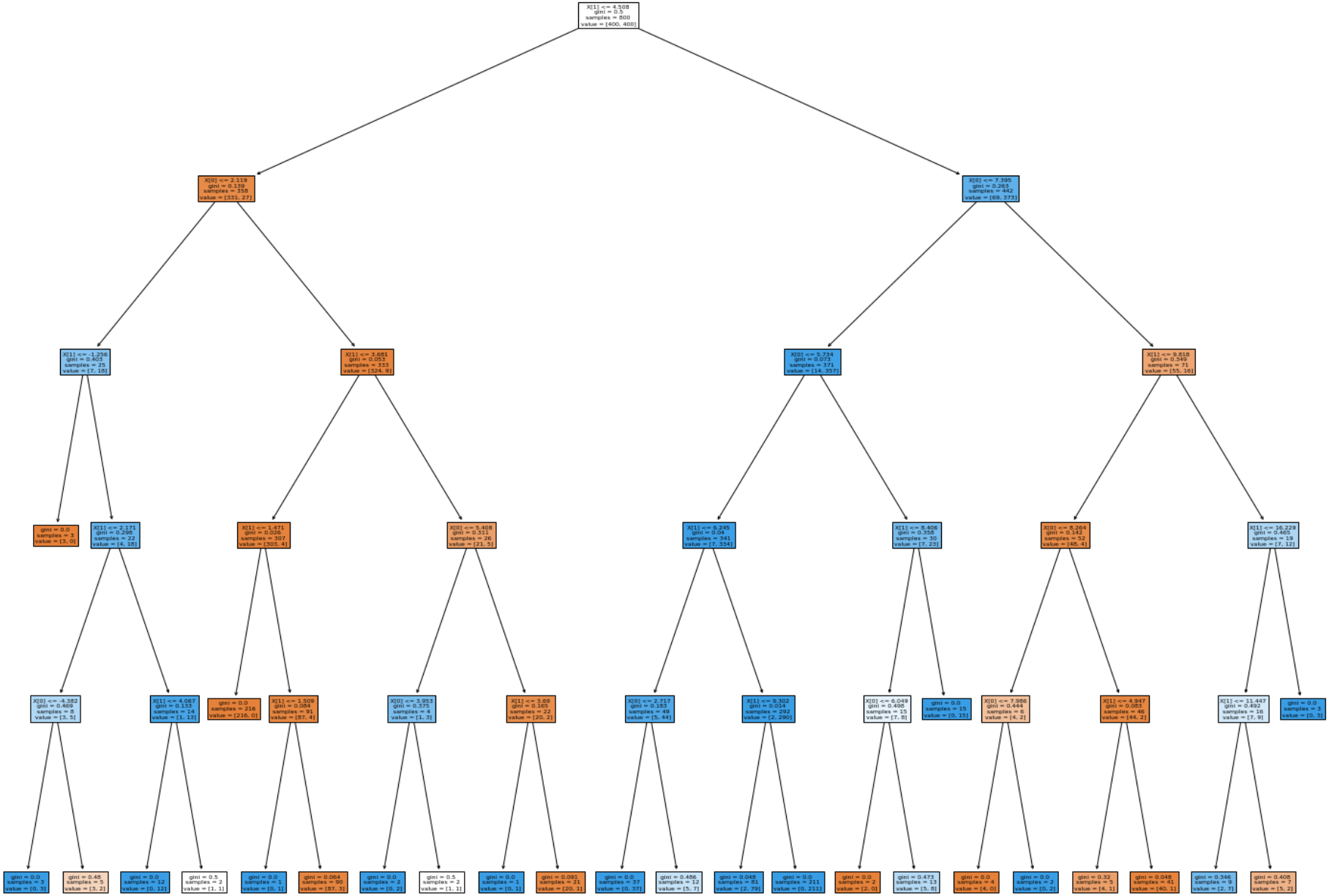

X[1] <= 4.508
gini = 0.5
samples = 800
value = [400, 400]
X[0] <= 2.119
gini = 0.139
samples = 358
value = [331, 27]
X[0] <= 7.395
gini = 0.263
samples = 442
value = [69, 373]
X[1] <= -1.256
gini = 0.403
samples = 25
value = [7, 18]
X[1] <= 3.681
gini = 0.053
samples = 333
value = [324, 9]
X[0] <= 5.734
gini = 0.073
samples = 371
value = [14, 357]
X[1] <= 9.818
gini = 0.349
samples = 71
value = [55, 16]
gini = 0.0
samples = 3
value = [3, 0]
X[1] <= 2.171
gini = 0.298
samples = 22
value = [4, 18]
X[1] <= 1.471
gini = 0.026
samples = 307
value = [303, 4]
X[0] <= 5.408
gini = 0.311
samples = 26
value = [21, 5]
X[1] <= 6.245
gini = 0.04
samples = 341
value = [7, 334]
X[1] <= 8.406
gini = 0.358
samples = 30
value = [7, 23]
X[0] <= 8.264
gini = 0.142
samples = 52
value = [48, 4]
X[1] <= 16.229
gini = 0.465
samples = 19
value = [7, 12]
gini = 0.469
samples = 8
value = [3, 5]
gini = 0.133
samples = 14
value = [1, 13]
gini = 0.0
samples = 216
value = [216, 0]
gini = 0.084
samples = 91
value = [87, 4]
gini = 0.375
samples = 4
value = [1, 3]
gini = 0.165
samples = 22
value = [20, 2]
gini = 0.183
samples = 49
value = [5, 44]
gini = 0.014
samples = 292
value = [2, 290]
gini = 0.498
samples = 15
value = [7, 8]
gini = 0.0
samples = 15
value = [0, 15]
gini = 0.444
samples = 6
value = [4, 2]
gini = 0.083
samples = 46
value = [44, 2]
gini = 0.492
samples = 16
value = [7, 9]
gini = 0.0
samples = 3
value = [0, 3]


Feature Space - Arbol-4 - Training: Acc = 0.9550
1
2
x_1
x_0


Feature Space - Arbol-4 - Testing: Acc = 0.9250
1
2
x_1
x_0


Feature Space - Arbol-5 - Training: Acc = 0.9675
1
2
x_1
x_0


Feature Space - Arbol-5 - Testing: Acc = 0.8900
1
2
x_1
x_0


X[1] <= 4.508
gini = 0.5
samples = 800
value = [400, 400]
X[0] <= 2.119
gini = 0.139
samples = 358
value = [331, 27]
X[0] <= 7.395
gini = 0.263
samples = 442
value = [69, 373]
X[1] <= -1.256
gini = 0.463
samples = 25
value = [7, 18]
X[1] <= 3.681
gini = 0.053
samples = 333
value = [324, 9]
X[0] <= 5.734
gini = 0.075
samples = 371
value = [14, 357]
X[1] <= 9.818
gini = 0.349
samples = 71
value = [55, 16]
gini = 0.0
samples = 3
value = [3, 0]
X[1] <= 2.171
gini = 0.298
samples = 22
value = [4, 18]
X[1] <= 1.471
gini = 0.026
samples = 307
value = [303, 4]
X[0] <= 5.408
gini = 0.311
samples = 26
value = [21, 5]
X[1] <= 8.245
gini = 0.04
samples = 341
value = [7, 334]
X[1] <= 8.406
gini = 0.356
samples = 30
value = [7, 23]
X[0] <= 8.264
gini = 0.142
samples = 52
value = [45, 4]
X[1] <= 16.229
gini = 0.465
samples = 19
value = [7, 12]
X[0] <= -4.352
gini = 0.469
samples = 8
value = [3, 5]
X[1] <= 4.067
gini = 0.133
samples = 14
value = [1, 13]
gini = 0.0
samples = 216
value = [216, 0]
X[1] <= 1.509
gini = 0.054
samples = 91
value = [87, 4]
X[0] <= 5.953
gini = 0.375
samples = 4
value = [1, 3]
X[1] <= 3.69
gini = 0.165
samples = 22
value = [20, 2]
X[0] <= 2.717
gini = 0.163
samples = 49
value = [5, 44]
X[1] <= 9.302
gini = 0.014
samples = 292
value = [2, 290]
X[0] <= 6.049
gini = 0.498
samples = 15
value = [7, 8]
gini = 0.0
samples = 15
value = [0, 15]
X[0] <= 7.986
gini = 0.444
samples = 6
value = [4, 2]
X[1] <= 4.947
gini = 0.067
samples = 46
value = [44, 2]
X[1] <= 11.447
gini = 0.492
samples = 16
value = [7, 9]
gini = 0.0
samples = 3
value = [0, 3]
gini = 0.0
samples = 3
value = [0, 3]
gini = 0.48
samples = 5
value = [3, 2]
gini = 0.0
samples = 12
value = [0, 12]
gini = 0.5
samples = 2
value = [1, 1]
gini = 0.0
samples = 1
value = [0, 1]
gini = 0.064
samples = 90
value = [87, 3]
gini = 0.0
samples = 2
value = [0, 2]
gini = 0.5
samples = 2
value = [1, 1]
gini = 0.0
samples = 1
value = [0, 1]
gini = 0.095
samples = 21
value = [20, 1]
gini = 0.0
samples = 37
value = [0, 37]
gini = 0.486
samples = 12
value = [5, 7]
gini = 0.048
samples = 81
value = [2, 79]
gini = 0.0
samples = 211
value = [0, 211]
gini = 0.0
samples = 2
value = [2, 0]
gini = 0.473
samples = 13
value = [5, 8]
gini = 0.0
samples = 4
value = [4, 0]
gini = 0.0
samples = 2
value = [0, 2]
gini = 0.32
samples = 5
value = [4, 1]
gini = 0.048
samples = 41
value = [40, 1]
gini = 0.346
samples = 9
value = [2, 7]
gini = 0.408
samples = 7
value = [5, 2]

Random Forest

Random Forest

TRAINING:

for i = 1 to n

Escoger aleatoriamente un subconjunto de training

Entrenar un árbol de decisión A_i

Random Forest

TRAINING:

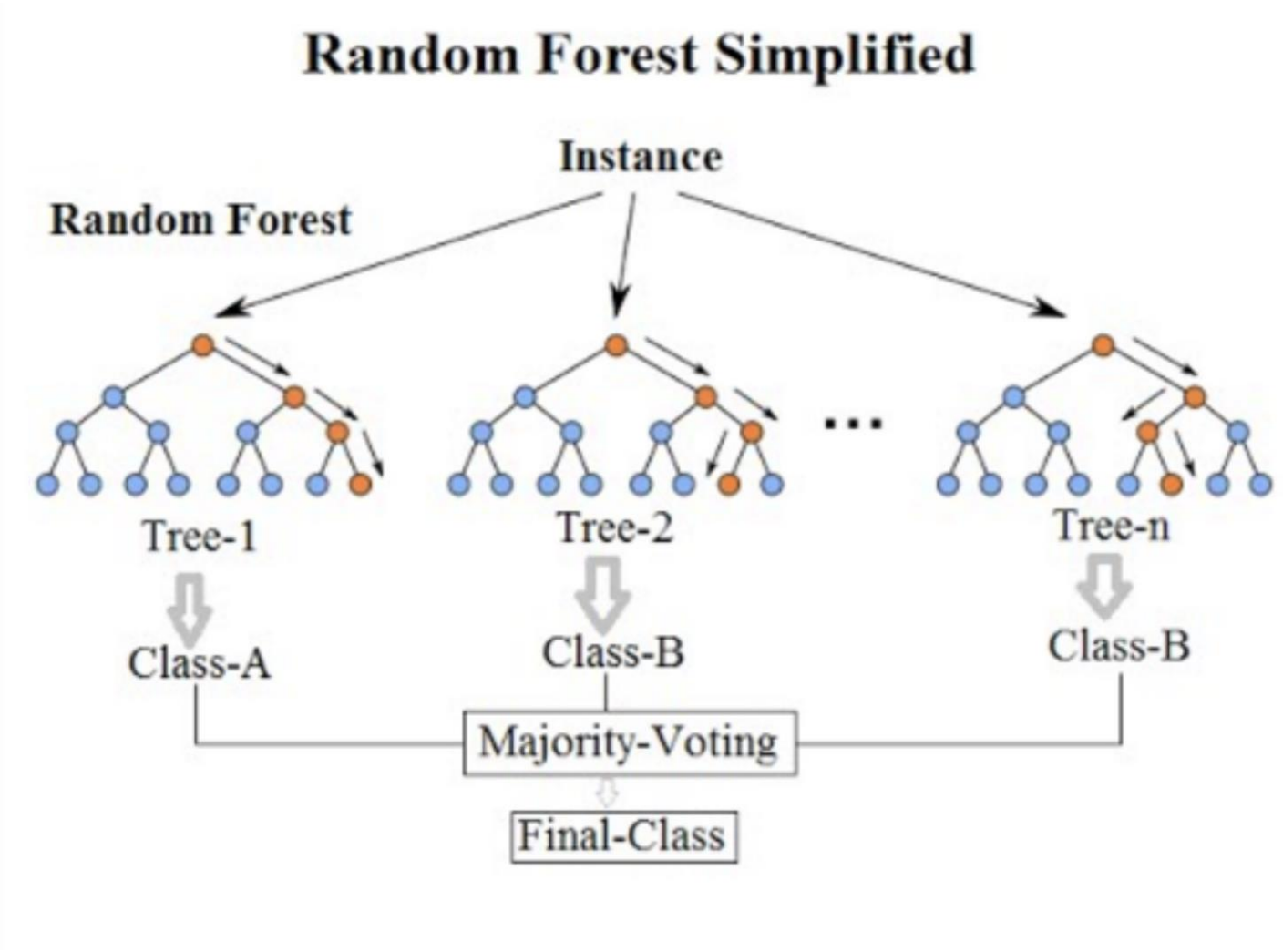
for i = 1 to n
    Escoger aleatoriamente un subconjunto de training
    Entrenar un árbol de decisión A_i

TESTING:

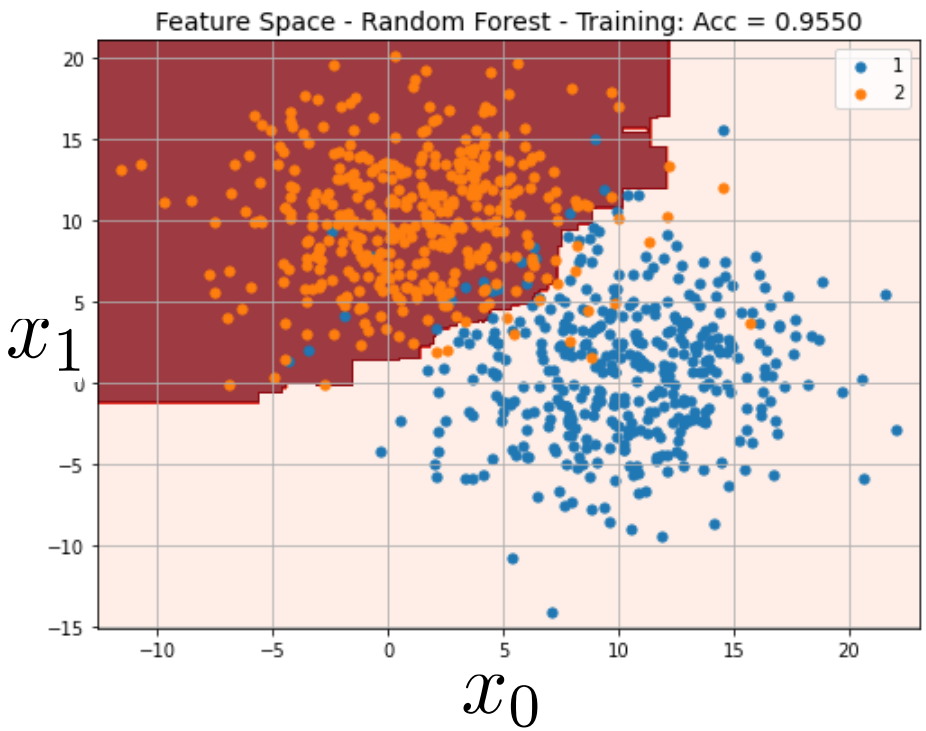
for i = 1 to n
    Clasificar la muestra de testing usando A_i

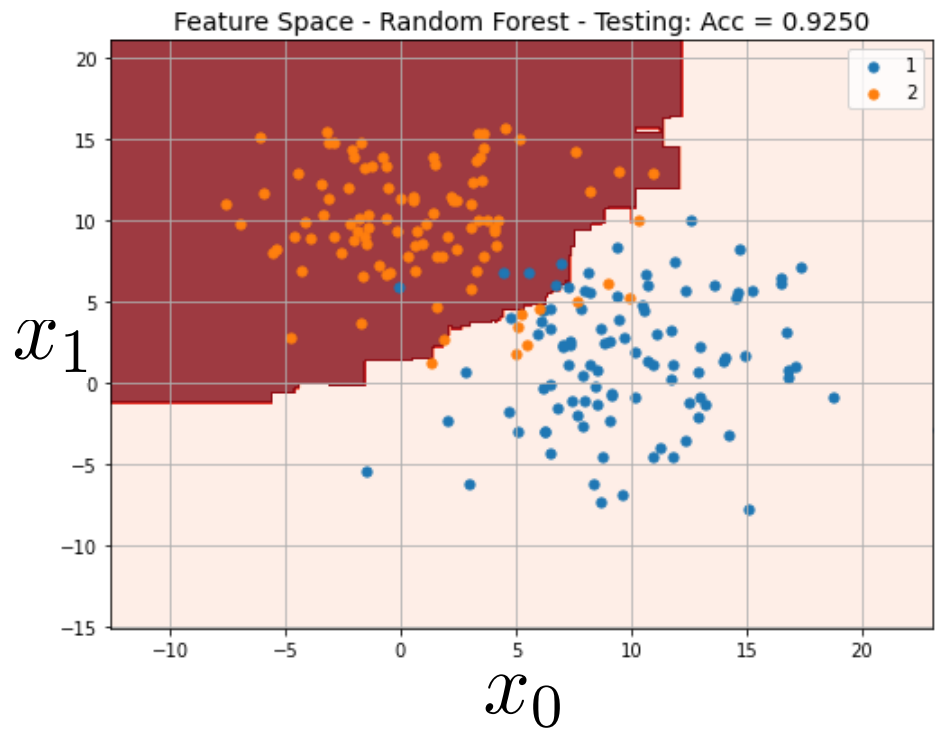
Clasificar la muestra según la mayoría de los n votos

Random Forest



https://en.wikipedia.org/wiki/Random_forest


Feature Space - Random Forest - Training: Acc = 0.9550
1
2
x_1
x_0


Feature Space - Random Forest - Testing: Acc = 0.9250
1
2
x_1
x_0